Las cuestiones prácticas sobre la causalidad han preocupado a los economistas durante varios siglos. Adam Smith escribió sobre las causas de la riqueza de las naciones (Smith, 2003). Karl Marx se interesó en la transición de la sociedad del capitalismo al socialismo (Needleman y Needleman, 1969). En el siglo XX, la Comisión Cowles buscó comprender mejor la identificación de los parámetros causales (Heckman y Vytlacil, 2007).[1] Los economistas han estado lidiando con las grandes ideas en torno a la causalidad y el desarrollo de herramientas empíricas útiles desde el principio.

Podemos observar el desarrollo de los conceptos modernos de causalidad en los escritos de varios filósofos. Hume (1993) describió la causalidad como una secuencia de eventos temporales en la que, si el primero no hubiera ocurrido, los posteriores tampoco lo habrían hecho. Por ejemplo, afirmó:

Podemos definir una causa como un objeto, seguido de otro, y donde todos los objetos similares al primero son seguidos por objetos similares al segundo. O, en otras palabras, donde, si el primer objeto no hubiera existido, el segundo nunca habría existido.

Mill (2010) ideó cinco métodos para inferir causalidad. Estos métodos fueron (1) el método de concordancia, (2) el método de la diferencia, (3) el método conjunto, (4) el método de variación concomitante y (5) el método de residuos. El segundo método, el método de la diferencia, es el más similar a la idea de causalidad como una comparación entre contrafácticos. Por ejemplo, escribió:

“Si una persona come de un plato en particular y muere como consecuencia, es decir, no habría muerto si no lo hubiera comido, la gente tendería a decir que comer ese plato fue la causa de su muerte”.

4.0.1 Inferencia estadística

Un avance significativo en nuestra comprensión de la causalidad coincide con el desarrollo de la estadística moderna. La teoría de la probabilidad y la estadística revolucionaron la ciencia en el siglo XIX, comenzando con la astronomía. Giuseppe Piazzi, astrónomo de principios del siglo XIX, descubrió el planeta enano Ceres, situado entre Júpiter y Marte, en 1801. Piazzi lo observó 24 veces antes de perderse de vista. Carl Friedrich Gauss propuso un método que permitía predecir con éxito la próxima ubicación de Ceres utilizando datos sobre su ubicación anterior. Su método minimizaba la suma de los errores al cuadrado; en otras palabras, el método de mínimos cuadrados ordinarios que analizamos anteriormente. Descubrió los MCO a los 18 años y publicó su derivación en 1809, a los 24 (Gauss, 1809).[2] Otros científicos que contribuyeron a nuestra comprensión de los MCO incluyen a Pierre-Simon LaPlace y Adrien-Marie Legendre.

El estadístico G. Udny Yule fue uno de los primeros en utilizar el análisis de regresión en las ciencias sociales. Yule (1899) se interesó en las causas de la pobreza en Inglaterra.

Las personas pobres dependían de las casas de beneficencia o de las autoridades locales para obtener apoyo financiero, y Yule quería saber si la asistencia pública aumentaba el número de indigentes, lo cual constituye una cuestión causal. Yule utilizó la regresión de mínimos cuadrados para estimar la correlación parcial entre la asistencia pública y la pobreza. Sus datos se extrajeron de los censos ingleses de 1871 y 1881, y los he puesto a disposición en mi sitio web para Stata o en la biblioteca Mixtape para usuarios de R. A continuación, se presenta un ejemplo de la regresión que se podría ejecutar con estos datos:

$$\text{Pauper} = \alpha + \delta\text{Outrelief} + \beta_1\text{Old} + \beta_2\text{Pop} + u$$

Let's run this regression using the data.

Stata Code | R Code | Python Code

yule.do

▼ Code

```
use https://github.com/scunning1975/mixtape/raw/master/yule.dta, clear
regress paup outrelief old pop
```

Cada fila de este conjunto de datos representa una ubicación específica en Inglaterra (p. ej., Chelsea, Strand). Por lo tanto, dado que hay 32 filas, el conjunto de datos contiene 32 ubicaciones en Inglaterra. Cada variable se expresa como una tasa de crecimiento anual. En consecuencia, cada coeficiente de regresión tiene interpretaciones de elasticidad, con una salvedad: técnicamente, como expliqué al principio del libro, las elasticidades son en realidad objetos causales, no simplemente correlaciones entre dos variables. Y es improbable que los datos de Yule cumplan las condiciones necesarias para interpretarlas como relaciones causales. No obstante, ejecutemos la regresión y observemos los resultados, que presento en la Tabla 4.1.

Table 4.1: Estimated association between pauperism growth rates and public assistance.

	Dependent variable
Covariates	**Pauperism growth**
Outrelief	0.752
	(0.135)
Old	0.056
	(0.223)
Pop	−0.311
	(0.067)

En otras palabras, un cambio de 10 puntos porcentuales en la tasa de crecimiento de la asistencia externa se asocia con un aumento de 7,5 puntos porcentuales en la tasa de crecimiento del pauperismo, o una elasticidad de 0,75. Yule utilizó su regresión para determinar la correlación entre la asistencia externa y el pauperismo, de la cual concluyó que la asistencia pública aumentaba las tasas de crecimiento de la pobreza.

Pero ¿qué podría estar mal en este razonamiento? ¿Hasta qué punto está convencido de que todas las vías indirectas entre el pauperismo y la asistencia externa se bloquean una vez que se controlan dos covariables en una base de datos transversal para toda Inglaterra? ¿Podría haber determinantes no observados tanto de la pobreza como de la asistencia pública? Después de todo, Yule no controla ningún factor económico, que sin duda afecta tanto a la pobreza como a la cantidad de recursos asignados a la asistencia externa. Asimismo, es posible que tenga la causalidad al revés: tal vez el aumento de la pobreza provoque que las comunidades aumenten la asistencia, y no solo al revés. Los primeros en adoptar una nueva metodología o técnica suelen ser quienes reciben más críticas, a pesar de ser pioneros de los propios métodos. Es facilísimo criticar duramente a un investigador de hace cien años, que trabajaba en una época en la que la alternativa a la regresión era una farsa ideológica. Además, no está aquí para responder. Simplemente quiero señalar que el uso ingenuo de la regresión para estimar correlaciones como forma de hacer afirmaciones causales que orienten importantes cuestiones políticas ha sido la norma durante mucho tiempo, y probablemente no desaparecerá pronto.

4.1 Aleatorización Física

La noción de aleatorización física como fundamento de la inferencia causal estuvo presente en los siglos XIX y XX, pero no fue hasta R. A. Fisher (1935) que se materializó. El primer experimento aleatorio históricamente reconocido había ocurrido cincuenta años antes en psicología (Peirce y Jastrow, 1885). Sin embargo, curiosamente, en ese experimento, la razón de la aleatorización no fue la base de la inferencia causal. Más

bien, los investigadores propusieron la aleatorización como una forma de engañar a los sujetos en sus experimentos. Peirce y Jastrow (1885) utilizaron varios tratamientos y emplearon la aleatorización física para que los participantes no pudieran adivinar qué sucedería a continuación. Salvo error, la recomendación de la aleatorización física de los tratamientos a las unidades como base de la inferencia causal se basa en Splawa-Neyman (1923) y Roland A. Fisher (1925). Más específicamente, Splawa-Neyman (1923) desarrolló la poderosa notación de resultados potenciales (que analizaremos pronto), y si bien propuso la aleatorización, esta no se consideró literalmente necesaria hasta Roland A. Fisher (1925). Roland A. Fisher (1925) propuso el uso explícito de la aleatorización en el diseño experimental para la inferencia causal.³

La aleatorización física se limitó principalmente a los experimentos agrícolas hasta mediados de la década de 1950, cuando comenzó a utilizarse en ensayos médicos. Entre los primeros experimentos aleatorizados importantes en medicina —de hecho, los que se intentaron— se encuentran los ensayos de campo de la vacuna antipoliomielítica de Salk. En 1954, el Servicio de Salud Pública se propuso determinar si la vacuna de Salk prevenía la polio. Los niños del estudio fueron asignados aleatoriamente para recibir la vacuna o un placebo.⁴ Además, los médicos que realizaban los diagnósticos de polio desconocían si el niño había recibido la vacuna o el placebo. El ensayo de la vacuna contra la polio se denominó ensayo controlado aleatorizado doble ciego porque ni el paciente ni quien administraba la vacuna sabían si el tratamiento era un placebo o una vacuna. Era necesario que el ensayo de campo fuera muy amplio, ya que la tasa de incidencia de polio en la población era de 50 por 100.000. El grupo de tratamiento, compuesto por 200.745 personas, presentó 33 casos de polio. El grupo de control, compuesto por 201.229 personas, presentó 115 casos. La probabilidad de observar una diferencia tan grande en las tasas de polio, solo por casualidad, es de aproximadamente una entre mil millones. Se argumentó que la única explicación plausible era que la vacuna contra la polio reducía el riesgo de contraerla.

Un experimento aleatorizado similar a gran escala se llevó a cabo en economía en la década de 1970. Entre 1971 y 1982, la Corporación RAND realizó un experimento aleatorizado a gran escala para estudiar el efecto causal del seguro médico en la utilización de la atención médica. Para el estudio, Rand reclutó a 7700 personas menores de 65 años. El experimento fue algo complejo, con múltiples grupos de tratamiento. Los participantes fueron asignados aleatoriamente a uno de cinco planes de seguro médico: atención gratuita, tres planes con diferentes niveles de costos compartidos y un plan HMO. Los participantes con costos compartidos realizaron menos visitas médicas y tuvieron menos hospitalizaciones que aquellos con atención gratuita. También se observaron otras disminuciones en la utilización de la atención médica, como menos visitas al dentista, entre los grupos de tratamiento con costos compartidos. En general, los participantes en los planes de costos compartidos tendieron a gastar menos en salud porque utilizaron menos servicios. La reducción en el uso de los servicios se debió principalmente a que los participantes en los grupos de tratamiento con costos compartidos optaron por no iniciar la atención.⁵

Sin embargo, el uso de experimentos aleatorios se ha disparado desde ese experimento de atención médica. Se han otorgado múltiples Premios Nobel a quienes los utilizan: Vernon Smith por ser pionero en los experimentos de laboratorio en 2002, y más recientemente, Abhijit Bannerjee, Esther Duflo y Michael Kremer en 2019 por aprovechar los experimentos de campo al servicio del alivio de la pobreza global.6 El diseño experimental se ha convertido en un sello distintivo en la microeconomía aplicada, la ciencia política, la sociología, la psicología y más. Pero ¿por qué se considera importante? ¿Por qué la aleatorización es un elemento tan clave de este diseño para aislar los efectos causales? Para comprender esto, necesitamos aprender más sobre la poderosa notación que Splawa-Neyman (1923) desarrolló, llamada "resultados potenciales".

4.1.1 Resultados potenciales

Si bien la notación de resultados potenciales se remonta a Splawa-Neyman (1923), adquirió un gran auge en las ciencias sociales en general con D. Rubin (1974).7 Al momento de escribir este libro, los resultados potenciales eran prácticamente la lingua franca para pensar y expresar enunciados causales, y probablemente le debemos a D. Rubin (1974) por ello tanto como cualquiera.

En la tradición de los resultados potenciales (Splawa-Neyman 1923; D. Rubin 1974), un efecto causal se define como una comparación entre dos estados del mundo. Permítanme ilustrarlo con un ejemplo sencillo. En el primer estado del mundo (a veces llamado el estado "real" del mundo), un hombre toma aspirina para su dolor de cabeza y una hora después informa la gravedad de su dolor. En el segundo estado del mundo (a veces llamado el estado "contrafactual"), ese mismo hombre no toma nada para su dolor de cabeza y una hora después informa la intensidad de este. ¿Cuál fue el efecto causal de la aspirina? Según la tradición de los resultados potenciales, el efecto causal de la aspirina es la diferencia en la intensidad de su dolor de cabeza entre dos estados del mundo: uno en el que tomó la aspirina (el estado real del mundo) y otro en el que nunca la tomó (el estado contrafactual). La diferencia en la intensidad del dolor de cabeza entre estos dos estados del mundo, medida en lo que por lo demás es el mismo momento, es el efecto causal de la aspirina en su dolor de cabeza. ¡Parece fácil!

Hacer preguntas como esta (y mucho menos intentar responderlas) es involucrarse en la narración. A los humanos siempre nos han interesado las historias que exploran contrafactuales. ¿Qué pasaría si los padres de Bruce Wayne nunca hubieran sido asesinados? ¿Y si esa camarera hubiera ganado la lotería? ¿Y si tu amigo del instituto nunca hubiera tomado esa primera copa? ¿Y si en Matrix, Neo hubiera tomado la pastilla azul? Estas hipótesis son divertidas, pero en última instancia, siguen siendo narrativas. Necesitamos que el Doctor Strange nos dé la Gema del Tiempo para responder a preguntas como estas.

Probablemente puedas entender adónde nos lleva esto. La notación de resultados potenciales expresa causalidad en términos de contrafácticos, y dado que estos no existen, la confianza en los efectos causales debe ser, hasta cierto punto, incontestable. Preguntarse cómo sería la vida si un solo evento hubiera sido diferente es caer en el razonamiento contrafáctico, y los contrafácticos no se materializan en la historia porque son estados hipotéticos del mundo. Por lo tanto, si la respuesta requiere datos sobre esos contrafácticos, entonces la pregunta no puede responderse. La historia es una secuencia de eventos observables y fácticos, uno tras otro. No sabemos qué habría sucedido si un evento hubiera cambiado porque carecemos de datos sobre el resultado contrafáctico.8 Los resultados potenciales existen ex ante como un conjunto de posibilidades, pero una vez que se toma una decisión, todos los resultados menos uno desaparecen.9

Para concretar esto, introduzcamos algo de notación y conceptos más específicos. Para simplificar, asumiremos una variable binaria que toma el valor 1 si una unidad en particular recibe el tratamiento y 0 si no.¹⁰ Cada unidad tendrá dos resultados potenciales, pero solo un resultado observado. Los resultados potenciales se definen como Y_i^1 Si la unidad i recibió el tratamiento, y como Yi0 si no lo recibió. Observe que ambos resultados potenciales tienen el mismo subíndice i; esto indica dos estados del mundo distintos para la misma persona en nuestro ejemplo, en el mismo momento. Llamaremos estado de control al estado del mundo donde no se realizó el tratamiento. Cada unidad i tiene exactamente dos resultados potenciales: un resultado potencial bajo un estado del mundo donde se realizó el tratamiento (Y1) y un resultado potencial donde no se realizó el tratamiento (Y0).

Los resultados observables o "reales", Yi, se distinguen de los resultados potenciales. En primer lugar, observe que los resultados reales no tienen superíndice. Esto se debe a que no son resultados potenciales, sino los resultados observados, reales, históricos, empíricos (como se quiera decir) que la unidad i experimentó. Mientras que los resultados potenciales son variables aleatorias hipotéticas que difieren en la población, los resultados observables son variables aleatorias factuales. Cómo pasamos de los resultados potenciales a los resultados reales es una importante cuestión filosófica, pero como cualquier buen economista, lo simplificaré con una ecuación. El resultado observable de una unidad es una función de sus resultados potenciales, determinados según la ecuación de conmutación:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Donde Di es igual a 1 si la unidad recibió el tratamiento y 0 si no. Observe la lógica de la ecuación. Cuando Di = 1, entonces Yi = Yi1, ya que el segundo término se anula. Y cuando Di = 0, el primer término se anula y, por lo tanto, Yi = Yi0. Utilizando esta notación, definimos el efecto del tratamiento específico de la unidad, o efecto causal, como la diferencia entre los dos estados del mundo:

$$\delta_i = Y_i^1 - Y_i^0$$

Inmediatamente nos enfrentamos a un problema. Si el efecto de un tratamiento requiere conocer dos estados del mundo, Yi1 e Yi0, pero mediante la ecuación de conmutación solo observamos uno, entonces no podemos calcular el efecto del tratamiento. Aquí radica el problema fundamental de la inferencia causal: la certeza en torno a los efectos causales requiere acceso a datos que siempre estarán disponibles.

4.1.2 Efectos promedio del tratamiento

De esta sencilla definición del efecto del tratamiento se desprenden tres parámetros diferentes que suelen ser de interés para los investigadores. Todos son medias poblacionales. El primero se denomina efecto promedio del tratamiento:

$$\begin{aligned} ATE &= E[\delta_i] \\ &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

Observe que, al igual que con nuestra definición de efectos del tratamiento a nivel individual, el efecto promedio del tratamiento requiere ambos resultados potenciales para cada unidad. Dado que solo conocemos uno de estos mediante la ecuación de conmutación, el efecto promedio del tratamiento, o ATE, es inherentemente desconocido. Por lo tanto, el ATE, al igual que el efecto individual del tratamiento, no es una cantidad que pueda calcularse. Sin embargo, sí puede estimarse.

El segundo parámetro de interés es el efecto promedio del tratamiento para el grupo de tratamiento. Es un nombre largo, pero permítanme explicarlo. Existen dos grupos de personas en esta discusión: un grupo de tratamiento y un grupo de control. El efecto promedio del tratamiento para el grupo de tratamiento, o ATT, es simplemente el efecto promedio del tratamiento poblacional para el grupo de unidades a las que se les asignó el tratamiento inicialmente según la ecuación de conmutación. En la medida en que

difiera en la población, es probable que el ATT difiera del ATE. En datos observacionales con seres humanos, casi siempre será diferente del ATE, debido a que los individuos optarán endógenamente por un tratamiento en función de los beneficios que esperan obtener de él. Al igual que el ATE, el ATT es incognoscible, ya que, al igual que este, también requiere dos observaciones por unidad de tratamiento. Formalmente, el ATT se expresa como:

$$\begin{aligned} ATT &= E[\delta_i \mid D_i = 1] \\ &= E[Y_i^1 - Y_i^0 \mid D_i = 1] \\ &= E[Y_i^1 \mid D_i = 1] - E[Y_i^0 \mid D_i = 1] \end{aligned}$$

El último parámetro de interés se denomina efecto promedio del tratamiento para el grupo de control o grupo sin tratamiento. Su abreviatura es ATU, que significa efecto promedio del tratamiento para los no tratados. Al igual que el ATT, el ATU es simplemente el efecto promedio del tratamiento poblacional para las unidades que se clasificaron en el grupo de control.¹¹ Dados los efectos heterogéneos del tratamiento, es probable que el ATT ≠ ATU, especialmente en un entorno observacional. La fórmula para el ATU es la siguiente.

$$
\begin{aligned}
ATU &= E[\delta_i \mid D_i = 0] \\
&= E[Y_i^1 - Y_i^0 \mid D_i = 0] \\
&= E[Y_i^1 \mid D_i = 0] - E[Y_i^0 \mid D_i = 0]
\end{aligned}
$$

Dependiendo de la pregunta de investigación, uno o los tres parámetros son interesantes. Sin embargo, los dos más comunes son la ATE y la ATT.

4.1.3 Descomposición simple por diferencia de medias

Esta discusión ha sido algo abstracta, así que seamos concretos. Supongamos que hay diez pacientes i con cáncer y dos procedimientos o tratamientos médicos. Hay una intervención quirúrgica, Di = 1, y hay una intervención de quimioterapia, Di = 0. Cada paciente tiene los siguientes dos resultados potenciales, donde un resultado potencial se define como la esperanza de vida después del tratamiento en años: un resultado potencial en un mundo donde se hubiera sometido a cirugía y un resultado potencial donde, en cambio, hubiera recibido quimioterapia. Usamos la notación Y1 e Y0, respectivamente, para estos dos estados del mundo.

Table 4.2: Potential outcomes for ten patients receiving surgery Y^1 or chemo Y^0.

Patients	Y^1	Y^0	δ
1	7	1	6
2	5	6	-1
3	5	1	4
4	7	8	-1
5	4	2	2
6	10	1	9
7	1	10	-9
8	5	6	-1
9	3	7	-4
10	9	8	1

Podemos calcular el efecto promedio del tratamiento si contamos con esta matriz de datos, ya que este es simplemente la diferencia media entre las columnas 2 y 3. Es decir, E[Y1]=5,6 y E[Y0]=5, lo que significa que ATE=0,6. En otras palabras, el efecto promedio del tratamiento quirúrgico en estos pacientes específicos es de 0,6 años adicionales (en comparación con la quimioterapia).

Pero esto es solo el promedio. Cabe destacar que no todos se benefician de la cirugía. El paciente 7, por ejemplo, vive solo un año más después de la cirugía, en comparación con diez años adicionales después de la quimioterapia. Pero el ATE es simplemente el promedio de estos efectos heterogéneos del tratamiento.

Para mantener esta ficción, supongamos que existe un médico ideal que conoce los resultados potenciales de cada persona y elige el tratamiento que maximiza su esperanza de vida después del tratamiento.¹² En otras palabras, el médico decide someter a un paciente a cirugía o quimioterapia dependiendo del tratamiento que tenga la mayor esperanza de vida después del tratamiento. Una vez que realiza la asignación del tratamiento, el médico observa el resultado real del tratamiento de acuerdo con la ecuación de cambio mencionada anteriormente.

Table 4.3: Post-treatment observed lifespans in years for surgery $D = 1$ versus chemotherapy $D = 0$.

Patients	Y	D
1	7	1
2	6	0
3	5	1
4	8	0
5	4	1
6	10	1
7	10	0
8	6	0
9	7	0
10	9	1

La Tabla 4.3 muestra únicamente el resultado observado para los grupos de tratamiento y control. La Tabla 4.3 difiere de la Tabla 4.2, que muestra los resultados potenciales de cada unidad. Una vez asignado el tratamiento, podemos calcular el efecto promedio del tratamiento para el grupo de cirugía (ATT) frente al grupo de quimioterapia (ATU). El ATT es

de 4,4 y el ATU es de -3,2. Esto significa que la esperanza de vida posoperatoria promedio para el grupo de cirugía es de 4,4 años adicionales, mientras que para el grupo de quimioterapia es de 3,2 años menos.¹³

Ahora, el ATE es de 0,6, que es simplemente un promedio ponderado entre el ATT y el ATU.¹⁴ Por lo tanto, sabemos que el efecto general de la cirugía es positivo, aunque el efecto para algunos es negativo. En otras palabras, existen efectos de tratamiento heterogéneos, pero el efecto neto es positivo. ¿Qué sucedería si simplemente comparáramos la esperanza de vida posoperatoria promedio de los dos grupos? Este estimador simplista se llama diferencia simple de medias y es una estimación de la ATE igual a

$$E[\,Y^1 \mid D = 1\,] - E[\,Y^0 \mid D = 0\,]$$

que puede estimarse utilizando muestras de datos:

$$\begin{aligned} SDO &= E[Y^1 \mid D = 1] - E[Y^0 \mid D = 0] \\ &= \frac{1}{N_T}\sum_{i=1}^{n}(\,y_i \mid d_i = 1) - \frac{1}{N_C}\sum_{i=1}^{n}(\,y_i \mid d_i = 0) \end{aligned}$$

Lo cual en esta situación es igual a 7−7,4 = −0,4. Esto significa que el grupo de tratamiento vive 0,4 años menos después de la cirugía que el grupo de quimioterapia cuando el médico perfecto asignó a cada unidad a su mejor tratamiento. Si bien la estadística es cierta, observe lo engañosa que es. Esta estadística, sin la debida calificación, podría fácilmente usarse para afirmar que, en promedio, la cirugía es perjudicial, cuando sabemos que eso no es cierto. Es sesgada porque las unidades individuales se clasificaron óptimamente en su mejor opción de tratamiento, creando diferencias fundamentales entre el grupo de tratamiento y el grupo de control que son una función directa de los propios resultados potenciales. Para aclarar esto lo más posible, descompondremos la simple diferencia de medias en tres partes. Esas tres partes se enumeran a continuación:

$$\begin{aligned} E[\,Y^1 \mid D = 1\,] - E[\,Y^0 \mid D = 0\,] &= ATE \\ &+ E[\,Y^0 \mid D = 1\,] - E[\,Y^0 \mid D = 0\,] \\ &+ (1-\pi)(ATT - ATU) \end{aligned}$$

Para comprender el origen de estas partes del lado derecho, debemos empezar de nuevo y descomponer el parámetro de interés, ATE, en sus componentes básicos. ATE es igual a la suma ponderada de las expectativas promedio condicionales, ATT y ATU.

$$\begin{aligned} ATE &= \pi ATT + (1-\pi)ATU \\ &= \pi E[\,Y^1 \mid D = 1\,] - \pi E[\,Y^0 \mid D = 1\,] \\ &+ (1-\pi)E[\,Y^1 \mid D = 0\,] - (1-\pi)E[\,Y^0 \mid D = 0\,] \\ &= \{\pi E[\,Y^1 \mid D = 1\,] + (1-\pi)E[\,Y^1 \mid D = 0\,]\} \\ &- \{\pi E[\,Y^0 \mid D = 1\,] + (1-\pi)E[\,Y^0 \mid D = 0\,]\} \end{aligned}$$

Donde π es la proporción de pacientes que se sometieron a cirugía y 1−π es la proporción de pacientes que recibieron quimioterapia. Dado que la notación de esperanza condicional es algo engorrosa, intercambiemos cada término del lado izquierdo, ATE, y del lado derecho con algunas letras. Esto simplificará la demostración:

$$\begin{aligned} E[Y^1 \mid D=1] &= a \\ E[Y^1 \mid D=0] &= b \\ E[Y^0 \mid D=1] &= c \\ E[Y^0 \mid D=0] &= d \\ ATE &= e \end{aligned}$$

Ahora que hemos hecho estas sustituciones, reorganicemos las letras redefiniendo ATE como un promedio ponderado de todas las expectativas condicionales.

$$\begin{aligned} e &= \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\} \\ e &= \pi a + b - \pi b - \pi c - d + \pi d \\ e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a}-\mathbf{a}) + (\mathbf{c}-\mathbf{c}) + (\mathbf{d}-\mathbf{d}) \\ 0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\ \mathbf{a}-\mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\ \mathbf{a}-\mathbf{d} &= e + (\mathbf{c}-\mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\ \mathbf{a}-\mathbf{d} &= e + (\mathbf{c}-\mathbf{d}) + (1-\pi)a - (1-\pi)b + (1-\pi)d - (1-\pi)c \\ \mathbf{a}-\mathbf{d} &= e + (\mathbf{c}-\mathbf{d}) + (1-\pi)(a-c) - (1-\pi)(b-d) \end{aligned}$$

Ahora, sustituyendo nuestras definiciones, obtenemos lo siguiente:

$$\begin{aligned} E[Y^1 \mid D=1] - E[Y^0 \mid D=0] &= ATE \\ &+ (E[Y^0 \mid D=1] - E[Y^0 \mid D=0]) \\ &+ (1-\pi)(ATT - ATU) \end{aligned}$$

Y termina la descomposición. Ahora viene la parte divertida: ¡pensemos en lo que acabamos de crear! El lado izquierdo se puede estimar con una muestra de datos, ya que ambos resultados potenciales se convierten en resultados reales bajo la ecuación de conmutación. Esa es simplemente la diferencia simple en los resultados medios. El lado derecho es más interesante porque nos dice cuál es la diferencia simple en los resultados medios por definición. Vamos a etiquetarlo.

$$\underbrace{\frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i = 1) - \frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i = 0)}_{\text{Simple Difference in Outcomes}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}}$$

$$+ \underbrace{E[Y^0 \mid D=1] - E[Y^0 \mid D=0]}_{\text{Selection bias}}$$

$$+ \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogeneous treatment effect bias}}$$

Analicemos cada uno de estos términos por separado. El lado izquierdo es la diferencia simple en los resultados medios, y ya sabemos que es igual a -0,4. Dado que se trata de una descomposición, el lado derecho también debe ser igual a -0,4.

El primer término es el efecto promedio del tratamiento, que es el parámetro de interés, y sabemos que es igual a 0,6. Por lo tanto, los dos términos restantes deben ser la fuente del sesgo que causa que la diferencia simple en las medias sea negativa.

El segundo término se denomina sesgo de selección, que merece ser analizado. En este caso, el sesgo de selección es la diferencia inherente entre los dos grupos si ambos recibieron quimioterapia. Sin embargo, por lo general, es solo una descripción de las diferencias entre los dos grupos si nunca se hubiera realizado un tratamiento. En otras palabras, hay dos grupos: un grupo de cirugía y un grupo de quimioterapia. ¿Cómo difieren sus resultados potenciales bajo control? Observe que el primero es un contrafactual, mientras que el segundo es un resultado observado según la ecuación de conmutación. Podemos calcular esta diferencia aquí porque tenemos los resultados potenciales completos en la Tabla 4.2. Esa diferencia es igual a -4,8.

El tercer término es una forma de sesgo menos conocida, pero interesante. Además, si el enfoque es el ATE, entonces siempre está presente.¹⁵ El sesgo del efecto del tratamiento heterogéneo es simplemente la diferencia en la rentabilidad de la cirugía para los dos grupos multiplicada por la proporción de la población que está en el grupo de quimioterapia. Este último término es 0,5 × (4,4 − (−3,2)) o 3,8. Nótese, por si no es obvio, que la razón por la que π = 0,5 es porque 5 de cada 10 unidades están en el grupo de quimioterapia.

Ahora que tenemos los tres parámetros a la derecha, podemos ver por qué la diferencia simple en los resultados medios es igual a -0,4.

$$-0.4 = 0.6 - 4.8 + 3.8$$

Lo que encuentro interesante, e incluso esperanzador, de esta descomposición es que muestra que un contraste entre el grupo de tratamiento y el grupo de control técnicamente "contiene" el parámetro de interés. Puse "contiene" entre comillas porque, si bien es claramente visible en la descomposición, la simple diferencia en los resultados no se presenta como la suma de tres partes. Más bien, la simple diferencia en los resultados no es más que un número. El número es la suma de las tres partes, pero no podemos calcular cada parte individualmente porque no disponemos de datos sobre los resultados contrafácticos subyacentes necesarios para realizar los cálculos. El problema es que ese parámetro de interés ha sido enmascarado por dos tipos de sesgo: el sesgo de selección y el sesgo del efecto heterogéneo del tratamiento. Si los conociéramos, podríamos simplemente restarlos, pero normalmente no los conocemos. Desarrollamos estrategias para contrarrestar estos sesgos, pero no podemos calcularlos directamente, como tampoco podemos calcular directamente el ATE, ya que estos sesgos dependen de contra fácticos no observables.

El problema tampoco radica en asumir heterogeneidad. Podemos asumir firmemente que los efectos del tratamiento son constantes, $\delta_i = \delta \; \forall i$, lo que provocará que ATU = ATT y que SDO = ATE + sesgo de selección. Pero aún así, ese desagradable sesgo de selección seguiría arruinando las cosas. Se podría argumentar que toda la tarea de la inferencia

causal consiste en desarrollar una estrategia razonable para negar el papel que el sesgo de selección desempeña en los efectos causales estimados.

4.1.4 Supuesto de independencia

Comencemos con la situación más creíble para usar el SDO para estimar el ATE: cuando el tratamiento en sí (p. ej., cirugía) se ha asignado a los pacientes independientemente de sus posibles resultados. Pero, ¿qué significa la palabra "independencia"? Bueno, en términos de notación, significa:

$$(Y^1, Y^0) \perp D$$

Esto significa que se asignó cirugía a un individuo por razones ajenas a las ganancias de la cirugía.¹⁶ En nuestro ejemplo, ya sabemos que esto se viola porque el médico perfecto eligió específicamente cirugía o quimioterapia basándose en los resultados potenciales. Específicamente, un paciente recibió cirugía si Y1 > Y0 y quimioterapia si Y1 < Y0. Por lo tanto, en nuestro caso, el médico perfecto se aseguró de que D dependiera de Y1 e Y0. Todas las formas de clasificación basadas en humanos —probablemente como regla, para ser honestos— violan la independencia, que es la razón principal por la que las comparaciones observacionales ingenuas son casi siempre incapaces de recuperar efectos causales.¹⁷

Pero ¿y si no hubiera hecho eso? ¿Y si hubiera elegido la cirugía de tal manera que no dependiera de Y1 o Y0? ¿Cómo se elige la cirugía independientemente de las ganancias esperadas de la cirugía? Por ejemplo, tal vez los ordenó alfabéticamente por apellido, y los primeros cinco recibieron cirugía y los últimos cinco recibieron quimioterapia. O tal vez usó el segundero de su reloj para asignarles la cirugía: si era entre 1 y 30 segundos, les administraba la cirugía, y si era entre 31 y 60 segundos, les administraba quimioterapia.¹⁸ En otras palabras, supongamos que eligió un método para asignar el tratamiento que no dependiera de los valores de los resultados potenciales en ninguno de los dos estados del mundo. ¿Qué significaría eso en este contexto? Bueno, significaría:

$$E[\,Y^1 \mid D = 1\,] - E[\,Y^1 \mid D = 0\,] = 0$$
$$E[\,Y^0 \mid D = 1\,] - E[\,Y^0 \mid D = 0\,] = 0$$

En otras palabras, significaría que el resultado potencial medio para

O es el mismo (en la población) tanto para el grupo de cirugía como para el de quimioterapia. Este tipo de aleatorización de la asignación del tratamiento eliminaría tanto el sesgo de selección como el sesgo de efecto heterogéneo del tratamiento. Analicémoslo en orden. El sesgo de selección se anula de la siguiente manera:

$$E[\,Y^0 \mid D = 1\,] - E[\,Y^0 \mid D = 0\,] = 0$$

Y, por lo tanto, el

ya no sufre sesgo de selección. ¿Cómo afecta la aleatorización al sesgo de tratamiento por heterogeneidad desde la tercera línea? Reescriba las definiciones de ATT y ATU:

$$ATT = E[Y^1 \mid D = 1] - E[Y^0 \mid D = 1]$$
$$ATU = E[Y^1 \mid D = 0] - E[Y^0 \mid D = 0]$$

Reescribe el sesgo de la tercera fila después de 1−π:

$$\begin{aligned} ATT - ATU \quad &= \mathbf{E}\left[\mathbf{Y^1} \mid \mathbf{D = 1}\right] - E[Y^0 \mid D = 1] \\ &\quad -\mathbf{E}\left[\mathbf{Y^1} \mid \mathbf{D = 0}\right] + E[Y^0 \mid D = 0] \\ &= 0 \end{aligned}$$

Si el tratamiento es independiente de los resultados potenciales, entonces:

$$\begin{aligned} \underbrace{\frac{1}{N_T}\sum_{i=1}^{n}(y_i \mid d_i = 1) - \frac{1}{N_C}\sum_{i=1}^{n}(y_i \mid d_i = 0)}_{SDO} \quad &= E[Y^1] - E[Y^0] \\ &= ATE \end{aligned}$$

En esta situación, lo que se necesita es simplemente (a) datos sobre los resultados observables, (b) datos sobre la asignación del tratamiento y (c) (Y1,Y0)⊥⊥D. Llamamos (c) el supuesto de independencia. Para ilustrar que esto conduciría al SDO, utilizamos la siguiente simulación de Monte Carlo. Nótese que el ATE en este ejemplo es igual a 0,6.

Codigo stata

```
clear all

program define gap, rclass


  version 14.2

  syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]

  clear

  drop _all

  set obs 10

  gen    y1 = 7 in 1

  replace y1 = 5 in 2

  replace y1 = 5 in 3

  replace y1 = 7 in 4

  replace y1 = 4 in 5

  replace y1 = 10 in 6

  replace y1 = 1 in 7
```

```
    replace y1 = 5 in 8
    replace y1 = 3 in 9
    replace y1 = 9 in 10

    gen     y0 = 1 in 1
    replace y0 = 6 in 2
    replace y0 = 1 in 3
    replace y0 = 8 in 4
    replace y0 = 2 in 5
    replace y0 = 1 in 6
    replace y0 = 10 in 7
    replace y0 = 6 in 8
    replace y0 = 7 in 9
    replace y0 = 8 in 10
    drawnorm random
    sort random

    gen     d=1 in 1/5
    replace d=0 in 6/10
    gen     y=d*y1 + (1-d)*y0
    egen sy1 = mean(y) if d==1
    egen sy0 = mean(y) if d==0
    collapse (mean) sy1 sy0
    gen sdo = sy1 - sy0
    keep sdo
    summarize sdo
    gen mean = r(mean)
    end

simulate mean, reps(10000): gap
su _sim_1
```

Este Monte Carlo se ejecuta 10 000 veces, calculando cada vez el SDO promedio bajo independencia, lo cual se garantiza mediante la ordenación aleatoria de números. En mi ejecución de este programa, el ATE es 0,6 y el SDO es, en promedio, igual a 0,59088,19.

Antes de pasar al SDO, enfaticemos algo que a menudo se pasa por alto en los estudiantes que aprenden por primera vez el concepto y la notación de independencia. La independencia no implica que $E[Y1|D=1]-E[Y0|D=0]=0$. Tampoco implica que $E[Y1|D=1]-E[Y0|D=1]=0$. Más bien, implica $E[Y1|D=1]-E[Y1|D=0]=0$ en una población grande.[2]0 Es decir, la independencia implica que los dos grupos de unidades, cirugía y quimioterapia, tienen el mismo resultado potencial en promedio en la población. ¿Cuán realista es la independencia en los datos observacionales? La economía, quizás más que cualquier otra ciencia, nos dice que es improbable que la independencia se mantenga observacionalmente. Los actores económicos siempre intentan alcanzar un óptimo. Por ejemplo, los padres asignan a sus hijos a la escuela que perciben como la mejor para ellos, basándose en los resultados potenciales. En otras palabras, las personas eligen sus intervenciones y, muy probablemente, sus decisiones están relacionadas con los resultados potenciales, lo que hace que las comparaciones simples sean inapropiadas. La elección racional siempre contradice el supuesto de independencia y, por lo tanto, la simple comparación de medias no se aproximará al verdadero efecto causal. Necesitamos la aleatorización unitaria para las comparaciones simples a fin de comprender los efectos causales en juego.

4.1.5 SUTVA

Rubin argumenta que este tipo de cálculo se basa en un conjunto de supuestos, a los que denomina supuesto de valor de tratamiento unitario estable (SUTVA). Suena complejo, pero significa lo siguiente: nuestro marco de resultados potenciales nos impone límites para calcular los efectos del tratamiento. Cuando estos límites no se cumplen de forma creíble en los datos, debemos buscar una nueva solución. Estas limitaciones son que cada unidad recibe la misma dosis, no hay externalidades que afecten a los resultados potenciales de otras unidades cuando una unidad se expone a algún tratamiento, y no hay efectos de equilibrio general.

En primer lugar, esto implica que el tratamiento se recibe en dosis homogéneas para todas las unidades. Sin embargo, es fácil imaginar incumplimientos de esta norma; por ejemplo, si algunos médicos son mejores cirujanos que otros. En cuyo caso, simplemente debemos tener cuidado con lo que definimos como tratamiento y lo que no.

En segundo lugar, esto implica que no hay externalidades, porque, por definición, una externalidad se extiende a otras unidades no tratadas. En otras palabras, si la unidad 1 recibe el tratamiento y existe alguna externalidad, la unidad 2 tendrá un valor de Y0 diferente al que tendría si la unidad 1 no lo hubiera recibido. Estamos suponiendo que no se tiene en cuenta este tipo de efecto de contagio. Sin embargo, cuando existen tales efectos de contagio, como cuando trabajamos con datos de redes sociales, necesitaremos utilizar modelos que puedan explicar explícitamente dichas violaciones del SUTVA, como el de Goldsmith-Pinkham e Imbens (2013).

Relacionado con este problema de efectos de contagio está el del equilibrio general. Supongamos que estamos estimando el efecto causal de los rendimientos de la educación. El aumento de la educación universitaria, en equilibrio general, causaría un cambio en los salarios relativos diferente al que ocurre en equilibrio parcial. Este tipo de

problema de escalamiento es una preocupación común cuando se considera la extrapolación del diseño experimental a la implementación a gran escala de una intervención en cierta población.

Replicando la "demanda de conocimiento del estado serológico respecto al VIH". Rebecca Thornton es una prolífica y creativa economista del desarrollo. Su investigación ha abarcado diversos temas en desarrollo y ha evaluado cuestiones cruciales sobre políticas óptimas para el VIH, la demanda de aprendizaje, la circuncisión, la educación y más. Algunos de estos artículos se han convertido en logros importantes. Meticulosa y cuidadosa, se ha convertido en una destacada experta en VIH en el África subsahariana. Me gustaría hablar sobre un ambicioso proyecto que emprendió como estudiante de posgrado en la zona rural de Malawi, sobre si los incentivos económicos inducían a las personas a conocer su estado serológico y el efecto dominó de dicho conocimiento en las conductas sexuales de riesgo posteriores (Thornton, 2008).

El estudio de Thornton surge en un contexto político donde se creía que las pruebas del VIH podían utilizarse para combatir la epidemia. La idea era simple: si las personas conocían su estado serológico respecto al VIH, tal vez enterarse de que estaban infectadas les impulsaría a tomar precauciones, reduciendo así la tasa de infección. Por ejemplo, podrían buscar tratamiento médico, prolongando así su vida y la calidad de vida que disfrutaban. Pero al conocer su estado serológico, tal vez descubrir que eran VIH positivos les haría disminuir las conductas de alto riesgo. De ser así, el aumento de las pruebas podría generar fricciones en la propia red sexual que frenarían la epidemia. Esta política era tan sensata que las suposiciones en las que se basaba no fueron cuestionadas hasta que Thornton (2008) realizó un ingenioso experimento de campo en la Malawi rural. Sus resultados fueron, como en muchos estudios, una mezcla de buenas y malas noticias. Intentar comprender la demanda del estado serológico respecto al VIH, o el efecto del estado serológico sobre los comportamientos de salud, es generalmente imposible sin un experimento. En la medida en que las personas opten de forma óptima por aprender sobre su tipo de VIH o por adoptar hábitos saludables, es improbable que el conocimiento sobre su estado serológico sea independiente de los posibles resultados. Casi con toda seguridad, son precisamente estos resultados potenciales los que influyen en las decisiones tanto de obtener esa información como de adoptar cualquier tipo de comportamiento de riesgo. Por lo tanto, se necesitaría un experimento de campo para comprobar las suposiciones subyacentes a esta política sensata de utilizar las pruebas para combatir la epidemia.

¿Cómo lo hizo? A los encuestados de la zona rural de Malawi se les ofreció una prueba de VIH gratuita a domicilio y se les asignó aleatoriamente un cupón o cupones de entre 1 y 3 dólares. Estos cupones se podían canjear al visitar un centro de asesoramiento y pruebas voluntarias (APV) cercano. La noticia más alentadora fue que los incentivos monetarios resultaron muy eficaces para que las personas solicitaran los resultados de las pruebas. En promedio, los encuestados que recibieron un cupón con valor en efectivo tenían el doble de probabilidades de acudir al centro de APV para obtener los resultados de sus pruebas, en comparación con quienes no recibieron compensación. ¿De qué cuantía era este incentivo? Bueno, el incentivo promedio en su experimento equivalía a aproximadamente un día de salario. Pero encontró un comportamiento positivo de búsqueda de estatus incluso con el incentivo más pequeño, que equivalía a tan solo una

décima parte del salario diario. Thornton demostró que incluso pequeños incentivos monetarios podían utilizarse para animar a las personas a conocer su tipo de VIH, lo cual tiene obvias implicaciones políticas.

La segunda parte del experimento echó por tierra cualquier optimismo derivado de sus primeros resultados. Varios meses después de entregar los incentivos económicos a los encuestados, Thornton realizó un seguimiento y los entrevistó sobre sus comportamientos de salud posteriores. Los encuestados también tuvieron la oportunidad de comprar condones. Utilizando su asignación aleatoria de incentivos para conocer el estado serológico respecto al VIH, pudo aislar el efecto causal del propio conocimiento sobre la compra de condones, su indicador de participación en relaciones sexuales de riesgo. Halló que, condicionado a conocer el estado serológico respecto al VIH a partir de los incentivos aleatorios, las personas VIH positivas aumentaron el uso del condón en comparación con aquellas que no habían conocido sus resultados, pero solo en la forma de comprar dos condones adicionales. Este estudio sugirió que algunos tipos de divulgación, como las pruebas puerta a puerta, pueden hacer que las personas conozcan su tipo de VIH, especialmente cuando se combinan con incentivos. Sin embargo, el simple hecho de haber sido incentivado para conocer su estado serológico respecto al VIH puede no llevar por sí mismo a las personas VIH positivas a reducir la participación en conductas sexuales de alto riesgo, como tener relaciones sexuales sin condón.

El experimento de Thorton fue más complejo de lo que puedo representar aquí, y además, me centraré ahora únicamente en el aspecto de la transferencia de efectivo, en forma de cupones. Sin embargo, me centraré exclusivamente en los resultados de sus incentivos. Pero antes de hacerlo, veamos sus hallazgos. La Tabla 4.4 muestra sus hallazgos.

Dado que su proyecto utiliza la asignación aleatoria de transferencias de efectivo para identificar el efecto causal en el aprendizaje, crea automáticamente una asignación de tratamiento que es independiente de los resultados potenciales en consideración. Sabemos esto aunque no podamos probarlo directamente (es decir, los resultados potenciales son invisibles) porque conocemos cómo funciona la ciencia. En otras palabras, la aleatorización, por diseño, asigna el tratamiento independientemente de los resultados potenciales. Por lo tanto, las simples diferencias de medias son suficientes para obtener estimaciones básicas de los efectos causales.

Pero Thornton estimará un modelo de regresión lineal con controles en lugar de usar una simple diferencia de medias por varias razones. En primer lugar, esto le permite incluir diversos controles que pueden reducir la varianza residual y, por lo tanto, mejorar la precisión de sus estimaciones. Esto es valioso porque, al mejorar la precisión, puede descartar una gama más amplia de efectos del tratamiento que técnicamente están contenidos en sus intervalos de confianza. Aunque probablemente en este caso esto no sea tan importante dado que, como veremos, sus errores estándar son minúsculos.

Table 4.4: Impact of Monetary Incentives and Distance on Learning HIV Results (Thornton 2008)

	1	2	3	4	5
Any incentive	0.431***	0.309***	0.219***	0.220***	0.219 ***
	(0.023)	(0.026)	(0.029)	(0.029)	(0.029)
Amount of incentive		0.091***	0.274***	0.274***	0.273***
		(0.012)	(0.036)	(0.035)	(0.036)
Amount of incentive^2			−0.063***	−0.063***	−0.063***
			(0.011)	(0.011)	(0.011)
HIV	−0.055*	−0.052	−0.05	−0.058*	−0.055*
	(0.031)	(0.032)	(0.032)	(0.031)	(0.031)
Distance (km)				−0.076***	
				(0.027)	
Distance^2				0.010**	
				(0.005)	
Controls	Yes	Yes	Yes	Yes	Yes
Sample size	2,812	2,812	2,812	2,812	2,812
Average attendance	0.69	0.69	0.69	0.69	0.69

Las columnas 1 a 5 representan coeficientes MCO; errores estándar robustos agrupados por aldea (para 119 aldeas) con efectos fijos por distrito entre paréntesis. Todas las especificaciones también incluyen un término para la edad al cuadrado. "Cualquier incentivo" indica si el encuestado recibió algún incentivo monetario distinto de cero. "VIH" indica ser VIH positivo. "Distancia promedio simulada" es la distancia promedio de los hogares de los encuestados a ubicaciones aleatorias simuladas de centros de resultados de VIH. La distancia se mide como una distancia esférica en línea recta desde el hogar del encuestado a un centro de APV asignado aleatoriamente desde coordenadas geoespaciales y se mide en kilómetros.

Significativamente diferente de cero con un nivel de confianza del 99 %.

Significativamente diferente de cero con un nivel de confianza del 95 %.

Significativamente diferente de cero con un nivel de confianza del 90 %.

Pero la inclusión de controles tiene otro valor. Por ejemplo, si la asignación fue condicional a observables, o si la asignación se realizó en diferentes momentos, entonces incluir estos controles (como los efectos fijos por distrito) es técnicamente necesario para aislar los propios efectos causales. Y finalmente, la regresión genera errores estándar significativos, y quizás solo por eso, deberíamos darle una oportunidad.²¹

Entonces, ¿qué encontró Thornton? Utiliza mínimos cuadrados como modelo principal, representado en las columnas 1-5. Los tamaños del efecto que encuentra podrían describirse como gigantescos. Dado que solo el 34% de los participantes del grupo de control acudieron a un centro para conocer su estado serológico respecto al VIH, es impresionante que recibir algún dinero causara un aumento de 43 puntos porcentuales en el conocimiento del estado serológico respecto al VIH. Los incentivos monetarios, incluso los más pequeños, son suficientes para animar a muchas personas a recopilar datos de salud.

Las columnas 2-5 también son interesantes, pero no las detallaré aquí. En resumen, la columna 2 incluye un control para el monto del incentivo, que osciló entre US$0 y US$3. Esto nos permite estimar el impacto lineal de cada dólar adicional en el aprendizaje, que es relativamente pronunciado. Las columnas 3 a 5 incluyen una ecuación cuadrática y, como resultado, vemos que, si bien cada dólar adicional aumenta el aprendizaje, lo hace solo a un ritmo decreciente. Las columnas 4 y 5 incluyen controles para la distancia al centro de VCT y, al igual que en otros estudios, la distancia en sí misma es una barrera para algunos tipos de atención médica (Lindo et al., 2019).

Thornton también presenta un gráfico simple de sus resultados, que muestra diagramas de caja con la media y los intervalos de confianza para los grupos de tratamiento y control. Como veremos continuamente a lo largo del libro, los mejores artículos que estiman efectos causales siempre resumen sus resultados principales con imágenes ingeniosas y efectivas, y este estudio no es la excepción. Como muestra esta figura, los efectos fueron enormes.

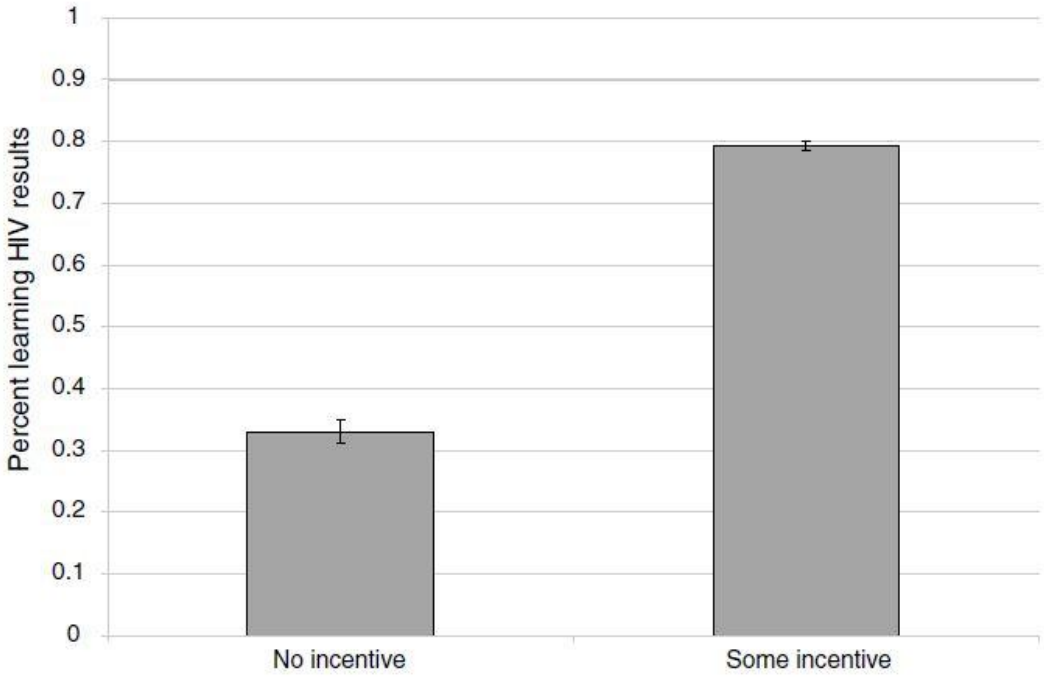


Figura 4.1: Representación visual de las transferencias de efectivo al conocer los resultados de la prueba del VIH (Thornton, 2008).

Si bien conocer el propio estado serológico respecto al VIH es importante, sobre todo si conlleva acceso a atención médica, los beneficios de las políticas que fomentan el aprendizaje son especialmente mayores si generan cambios en las conductas sexuales de alto riesgo entre las personas VIH positivas. De hecho, dados los efectos multiplicadores asociados con la introducción de fricciones en la red sexual mediante conductas de mitigación de riesgos (sobre todo si interrumpen las relaciones de pareja), estas iniciativas pueden ser tan beneficiosas que justifican muchos tipos de programas que, de otro modo, podrían no ser rentables.

Thornton examina en su encuesta de seguimiento, donde preguntó a todas las personas, independientemente de si conocían su estado serológico respecto al VIH, el efecto de una transferencia de efectivo en la compra de preservativos. Veamos primero sus principales resultados en la Figura 4.2.

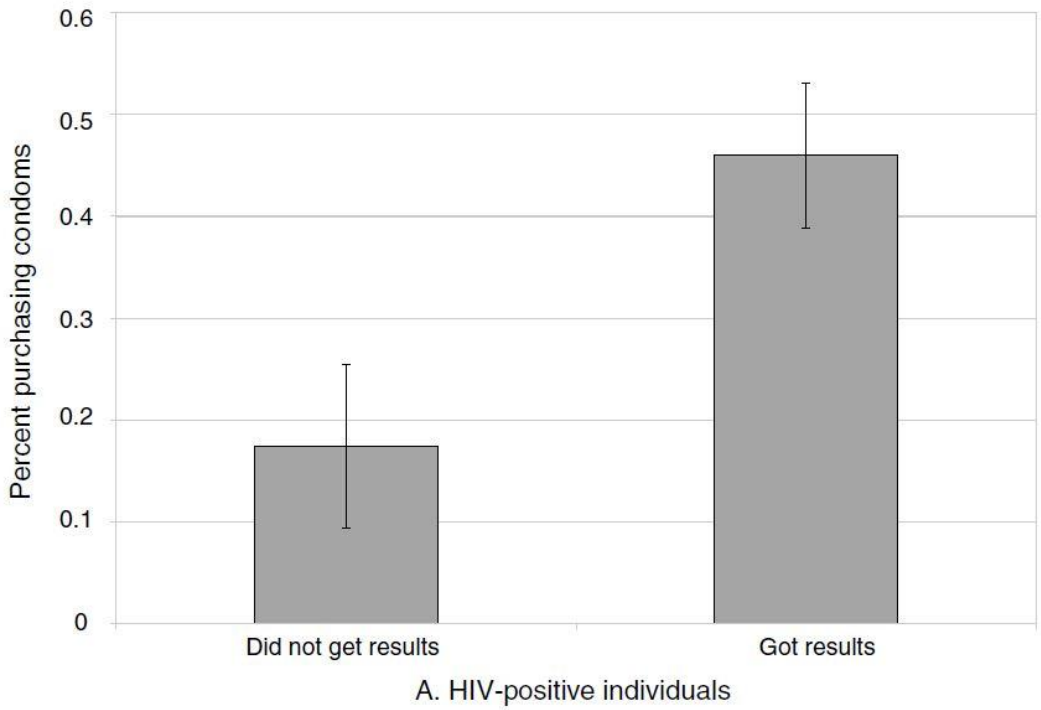


Figura 4.2: Representación visual de las transferencias de efectivo para la compra de preservativos para personas VIH positivas (Thornton, 2008).

Inicialmente, resulta alentador ver que los efectos en la compra de preservativos son considerables para las personas VIH positivas que, gracias al incentivo, obtuvieron los resultados de sus pruebas. El porcentaje de quienes compraron preservativos aumenta desde un punto de referencia de poco más del 30 % hasta un impresionante 80 % con cualquier incentivo. Pero lo desalentador es el número de preservativos adicionales que esto realmente implicó. En las columnas 3 y 4 de la Tabla 4.5, se observa el problema.

Figura 4.2: Representación visual de las transferencias de efectivo para la compra de preservativos para personas VIH positivas (Thornton, 2008).

Inicialmente, resulta alentador ver que los efectos en la compra de preservativos son considerables para las personas VIH positivas que, gracias al incentivo, obtuvieron los resultados de sus pruebas. El porcentaje de quienes compraron preservativos aumenta desde un punto de referencia de poco más del 30 % hasta un impresionante 80 % con cualquier incentivo. Pero lo desalentador es el número de preservativos adicionales que esto realmente implicó. En las columnas 3 y 4 de la Tabla 4.5, se observa el problema.

Table 4.5: Reactions to Learning HIV Results among Sexually Active at Baseline (Thornton 2008)

Dependent variables:				
Got results	−0.022	−0.069	−0.193	−0.303
	(0.025)	(0.062)	(0.148)	(0.285)
Got results × HIV	0.418***	0.248	1.778***	1.689**
	(0.143)	(0.169)	(0.564)	(0.784)
HIV	−0.175**	−0.073	−0.873	−0.831
	(0.085)	(0.123)	(0.275)	(0.375)
Controls	Yes	Yes	Yes	Yes
Sample size	1,008	1,008	1,008	1,008
Mean	0.26	0.26	0.95	0.95

Sample includes individuals who tested for HIV and have demographic data

Thornton, acertadamente, aborda la pregunta de dos maneras, tanto para el lector como para la precisión. Quiere conocer el efecto de obtener resultados, pero estos solo importan (1) para quienes obtuvieron su estado serológico y (2) para quienes eran VIH positivos. Los efectos no deberían importar si eran VIH negativos. Y, en última instancia, eso es lo que encuentra, pero ¿cómo va a responder a la primera? Aquí examina el efecto para quienes obtuvieron sus resultados y quienes eran VIH positivos mediante una interacción. Y esa es la columna 1: las personas que obtuvieron su estado serológico y supieron ser VIH positivos tuvieron un 41 % más de probabilidad de comprar condones varios meses después. Sin embargo, este resultado se reduce al utilizar la aleatorización de los incentivos en un marco de variables instrumentales, que analizaremos más adelante en el libro. El coeficiente se reduce casi a la mitad y sus intervalos de confianza son tan amplios que no podemos asegurar que los efectos sean inexistentes. Pero supongamos que la razón por la que no encontró un efecto en ningún comportamiento de compra es porque el tamaño de la muestra es tan pequeño que captar el efecto con IV es exigir demasiado a los datos. ¿Qué pasaría si usáramos algo con un poco más de

información, como el número de condones comprados? Y ahí es donde las cosas se vuelven pesimistas. Sí, Thornton encuentra evidencia de que las personas VIH positivas compraban más condones, pero al ver cuántos, descubrimos que solo fueron alrededor de 2 condones más en la visita de seguimiento (columnas 3-4). Y el efecto en el sexo en sí (no mostrado) fue negativo, pequeño (reducción del 4%) y no lo suficientemente preciso como para afirmarlo.

En conclusión, el estudio de Thornton es uno de esos estudios que encontramos con frecuencia en la inferencia causal, una mezcla de aspectos positivos y negativos. Es positivo en el sentido de que incentivar a las personas con pequeños incentivos las lleva a recopilar información sobre su propio estado serológico con respecto al VIH. Pero nuestro entusiasmo se desvanece cuando nos enteramos de que el efecto sobre los comportamientos de riesgo reales no es muy grande: apenas dos condones adicionales comprados varios meses después para los individuos VIH positivos probablemente no van a generar grandes externalidades positivas a menos que recaigan sobre los individuos VIH positivos con mayor riesgo.

4.2 Inferencia de aleatorización

Athey e Imbens (2017), en su capítulo sobre experimentos aleatorios, señalan que «en la inferencia basada en la aleatorización, la incertidumbre en las estimaciones surge naturalmente de la asignación aleatoria de los tratamientos, más que de un muestreo hipotético de una población amplia» (73). Athey e Imbens forman parte de una tendencia creciente de economistas que utilizan métodos basados en la aleatorización para inferir la probabilidad de que un coeficiente estimado no sea simplemente el resultado del cambio. Esta tendencia creciente utiliza métodos basados en la aleatorización para construir valores exactos que reflejan la probabilidad de que el azar haya producido la estimación.

¿Por qué se ha vuelto tan popular la inferencia basada en la aleatorización ahora? ¿Por qué no hace veinte años o más? No está claro por qué la inferencia basada en la aleatorización se ha vuelto tan popular en los últimos años, pero algunas posibilidades podrían explicar esta tendencia. Puede deberse al auge de los ensayos controlados aleatorios en economía, a la disponibilidad de bases de datos administrativas a gran escala que no son muestras de una población mayor, sino que representan "todos los datos", o a que la capacidad computacional ha mejorado tanto que la inferencia aleatoria se ha vuelto trivialmente sencilla de implementar al trabajar con miles de observaciones. Sea cual sea el motivo, la inferencia aleatoria se ha convertido en una forma muy común de referirse a la incertidumbre en torno a las estimaciones.

Existen al menos tres razones por las que podríamos realizar inferencia aleatoria. En primer lugar, puede deberse a que no trabajamos con muestras, y dado que los errores

estándar suelen justificarse alegando que reflejan la incertidumbre muestral, los métodos tradicionales podrían no ser tan significativos. La incertidumbre central en un estudio causal no se basa en la incertidumbre muestral, sino en el hecho de que desconocemos el contrafactual (Abadie, Diamond y Hainmueller, 2010; Abadie et al., 2020). En segundo lugar, puede que nos resulte incómodo apelar a las propiedades de un estimador en muestras grandes en un contexto particular, como cuando trabajamos con un número reducido de unidades de tratamiento. En tales situaciones, suponer que el número de unidades aumenta hasta el infinito podría poner en entredicho la credibilidad (Buchmueller, DiNardo y Valletta, 2011). Esto puede ser particularmente problemático en la práctica. Young (2019) muestra que, en muestras finitas, es común que algunas observaciones experimenten un apalancamiento concentrado. El apalancamiento provoca que los errores estándar y las estimaciones se vuelvan volátiles, lo que puede llevar a un rechazo excesivo. La inferencia aleatoria puede ser más robusta ante estos valores atípicos. Finalmente, parece haber cierta preferencia estética por este tipo de inferencia basada en placebo, ya que muchas personas la encuentran intuitiva. Si bien esto no es razón suficiente para adoptar un procedimiento metodológico, es muy común escuchar a alguien decir que utilizó la inferencia aleatoria porque tiene sentido. Pensé que valía la pena mencionarlo, ya que probablemente también se encontrará con comentarios similares. Pero antes de profundizar en ello, analicemos su historia, que se remonta a Ronald Fisher a principios del siglo XX.

4.2.1 Mujer degustando té

R. A. Fisher (1935) describió un experimento mental en el que una mujer afirma poder discernir si se vertió primero leche o té en una taza. Aunque no menciona su nombre, ahora sabemos que la mujer del experimento mental era Muriel Bristol y que, de hecho, el experimento mental se llevó a cabo.22 Muriel Bristol era una científica con doctorado en una época en la que las mujeres rara vez podían obtenerlo. Un día, durante el té de la tarde, Muriel afirmó que podía distinguir si la leche se añadía a la taza antes o después del té. Incrédula, Fisher ideó rápidamente un experimento para poner a prueba su supuesto talento.

La hipótesis, correctamente formulada, es que, dada una taza de té con leche, una mujer puede discernir si se añadió primero leche o té. Para comprobar su afirmación, se prepararon ocho tazas de té; en cuatro se añadió primero la leche y en cuatro, el té. ¿Cuántas tazas tiene que identificar correctamente para convencernos de su asombrosa habilidad? R. A. Fisher (1935) propuso un tipo de inferencia basada en permutaciones, un método que ahora llamamos prueba exacta de Fisher. La mujer posee la capacidad probabilísticamente, no con certeza, si la probabilidad de que adivine las cuatro correctamente es suficientemente baja. Hay $8 \times 7 \times 6 \times 5 = 1680$ maneras de elegir una primera taza, una segunda taza, una tercera taza y una cuarta taza, en orden. Hay $4 \times 3 \times 2 \times 1 = 24$ maneras de ordenar cuatro tazas. Por lo tanto, el número de maneras de elegir cuatro tazas de ocho es $168024 = 70$. Nótese que la mujer realiza el experimento seleccionando cuatro tazas. La probabilidad de que identifique correctamente las cuatro tazas es de 170, lo que equivale a $p = 0,014$.

Sin embargo, quizás estaría más convencido de este método si pudiera ver una simulación. Realicemos un ejercicio de combinación simple. Puede hacerlo con el siguiente código.

Codigo Stata

```
clear
capture log close

* Create the data. 4 cups with tea, 4 cups with milk.

set obs 8
gen cup = _n

* Assume she guesses the first cup (1), then the second cup (2), and so forth
gen    guess = 1 in 1
replace guess = 2 in 2
replace guess = 3 in 3
replace guess = 4 in 4
replace guess = 0 in 5
replace guess = 0 in 6
replace guess = 0 in 7
replace guess = 0 in 8
label variable guess "1: she guesses tea before milk then stops"

tempfile correct
save "`correct'", replace

* ssc install percom
combin cup, k(4)
gen permutation = _n
tempfile combo
save "`combo'", replace

destring cup*, replace
cross using `correct'
```

```
sort permutation cup


gen     correct = 0

replace correct = 1 if cup_1 == 1 & cup_2 == 2 & cup_3 == 3 & cup_4 == 4


* Calculation p-value

count if correct==1

local correct `r(N)'

count

local total `r(N)'

di `correct'/`total'

gen pvalue = (`correct')/(`total')

su pvalue


* pvalue equals 0.014


capture log close

exit
```

Observen que obtenemos la misma respuesta en ambos casos: 0,014. Volvamos a la Dra. Bristol. O bien no tiene la capacidad de discriminar el orden en que se sirvieron el té y la leche, y por lo tanto eligió las cuatro tazas correctas al azar, o bien (como ella misma dijo) sí tiene la capacidad de discriminar el orden en que se sirvieron los ingredientes de una bebida. Dado que elegir correctamente es altamente improbable (1 probabilidad entre 70), es razonable creer que posee el talento que siempre afirmó tener.

¿Qué hemos hecho exactamente? Bueno, lo que hemos hecho es proporcionar un valor de probabilidad exacto de que el fenómeno observado fue simplemente producto del azar. Nunca se puede ignorar el problema fundamental de la inferencia causal: nunca conocemos un efecto causal. Solo lo estimamos. Y luego nos basamos en otros procedimientos para obtener razones para creer que la cifra que calculamos probablemente sea un efecto causal. La inferencia aleatoria, como toda inferencia, es un andamiaje epistemológico para un tipo particular de creencia; específicamente, la probabilidad de que el azar haya creado este valor observado mediante un tipo particular de procedimiento. Pero este ejemplo, si bien motivó a Fisher a desarrollar este método, no es un diseño experimental en el que se estimen efectos causales. Por lo tanto, ahora

me gustaría ir más allá. Espero que el procedimiento de inferencia aleatoria se convierta en una herramienta más interesante y eficaz para formular afirmaciones causales creíbles.

4.2.2 Metodología del nulo agudo de Fisher

Analicemos con más detalle qué entendemos por inferencia aleatoria en un contexto más fácil de entender: un experimento literal o un cuasi-experimento. Concluiremos con código que ilustra cómo podríamos implementarla. La principal ventaja de la inferencia aleatoria es que nos permite realizar cálculos de probabilidad que revelan si los datos probablemente provienen de una distribución verdaderamente aleatoria.

La metodología no se puede entender sin comprender primero el concepto de la nula aguda de Fisher. Esta nula aguda de Fisher es una afirmación según la cual ninguna unidad de nuestros datos, al ser tratada, tuvo un efecto causal. Si bien es un concepto sutil y quizás no sea fácilmente comprensible, lo será mucho más una vez que analicemos algunos ejemplos. El valor de la nula aguda de Fisher reside en que nos permite realizar una inferencia "exacta" que no depende de distribuciones hipotéticas (p. ej., la gaussiana) ni de aproximaciones de muestras grandes. En este sentido, es no paramétrico.23

Algunos, al enfrentarse por primera vez al concepto de inferencia de aleatorización, piensan: «Esto suena a bootstrapping», pero en realidad son completamente diferentes. Los valores p de bootstrapping son extracciones aleatorias de la muestra que se utilizan para realizar la inferencia. Esto significa que el bootstrapping se centra principalmente en la incertidumbre de las observaciones utilizadas en la propia muestra. Sin embargo, los valores p de inferencia de aleatorización no se centran en la incertidumbre de la muestra; más bien, se basan en la incertidumbre sobre qué unidades de una muestra se asignan al tratamiento en sí.

Para ayudarle a comprender la inferencia de aleatorización, vamos a dividirla en algunos pasos metodológicos. Se podría decir que la inferencia de aleatorización consta de seis pasos: (1) la elección de la hipótesis nula precisa, (2) la construcción de la hipótesis nula, (3) la selección de un vector de tratamiento diferente, (4) el cálculo del estadístico de prueba correspondiente para ese nuevo vector de tratamiento, (5) la aleatorización a lo largo del paso 3, a medida que se recorren varios vectores de tratamiento nuevos (idealmente, todas las combinaciones posibles), y (6) el cálculo del valor p exacto.

4.2.3 Pasos para obtener un valor p

Fisher y Neyman debatieron sobre este primer paso. La hipótesis nula "precisa" de Fisher consistía en la afirmación de que cada unidad tenía un efecto del tratamiento de cero, lo que lleva a la sencilla afirmación de que el ATE también es cero. Neyman, por otro lado, partió de la otra dirección y afirmó que no había un efecto del tratamiento promedio, no que cada unidad tuviera un efecto del tratamiento cero. Esta es una distinción importante. Para comprender esto, supongamos que su efecto del tratamiento es 5, pero mi efecto del tratamiento es −5. Entonces, el ATE=0, que fue idea de Neyman. Pero la idea de Fisher era decir que mi efecto del tratamiento era cero, y el tuyo también. Esto es lo

que significa "agudo": significa literalmente que ninguna unidad individual tiene un efecto del tratamiento. Expresémoslo usando la notación de resultados potenciales, que puede ayudar a aclarar lo que quiero decir. H0:δi=Yi1−Yi0=0∀i. Puede que no sea obvio cómo esto nos ayudará, pero consideren lo siguiente: dado que conocemos todos los valores observados, si no hay efecto del tratamiento, también conocemos el contrafactual de cada unidad. Permítanme ilustrar mi punto con el ejemplo de la Tabla 4.6.

Table 4.6: Example of made-up data for 8 people with missing counterfactuals.

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

Si observa detenidamente la Tabla 15, verá que, para cada unidad, solo observamos un resultado potencial. Sin embargo, bajo la hipótesis nula aguda, podemos inferir el otro contrafáctico faltante. Solo disponemos de información sobre los resultados observados con base en la ecuación de conmutación. Por lo tanto, si una unidad recibe tratamiento, conocemos su Y1, pero no su Y0.

El segundo paso es la construcción de lo que se denomina un "estadístico de prueba". ¿Qué es esto? Un estadístico de prueba t(D,Y) es simplemente una cantidad escalar conocida que se calcula a partir de las asignaciones de tratamiento y los resultados observados. A menudo, no es más que una medida de la relación entre los valores de Y por D. En el resto de esta sección, desarrollaremos diversas maneras de construir estadísticos de prueba, pero comenzaremos con una medida bastante sencilla: la diferencia en el resultado medio.

En última instancia, los estadísticos de prueba nos ayudan a distinguir entre la hipótesis nula aguda en sí y alguna otra hipótesis. Y si se desea un estadístico de prueba con alta potencia estadística, se necesita que este adopte valores "extremos" (es decir, grandes en valores absolutos) cuando la hipótesis nula es falsa, y es necesario que estos valores grandes sean improbables cuando la hipótesis nula es verdadera.24

Como dijimos, existen varias maneras de estimar un estadístico de prueba, y analizaremos varias de ellas, pero comencemos con la simple diferencia en los resultados medios. Los valores promedio para el grupo de tratamiento son 34/4, los valores promedio para el grupo de control son 30/4, y la diferencia entre estos dos promedios es 1. Por lo tanto, dada esta asignación de tratamiento particular en nuestra muestra (la asignación verdadera, cabe mencionar), existe un estadístico de prueba correspondiente (la simple diferencia en los resultados medios) que es igual a 1.

Ahora bien, lo que implica la hipótesis nula aguda de Fisher es uno de los aspectos más interesantes de este método. Si bien históricamente desconocemos el contrafactual de cada unidad, bajo la hipótesis nula aguda sí conocemos el contrafactual de cada unidad. ¿Cómo es posible? Porque si ninguna de las unidades tiene efectos de tratamiento distintos de cero, entonces cada contrafáctico debe ser igual a su resultado observado. Esto significa que podemos completar los contrafácticos faltantes con los valores observados (Tabla 4.7).

Table 4.7: Example of made-up data for 8 people with filled-in counterfactuals according to Fisher's sharp null hypothesis.

Name	D	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Con estos contrafácticos faltantes reemplazados por el resultado observado correspondiente, no hay efecto del tratamiento a nivel de unidad y, por lo tanto, un ATE de cero. Entonces, ¿por qué encontramos antes una simple diferencia en los resultados medios de 1 si, de hecho, no hubo efecto promedio del tratamiento? Simple: era solo ruido, puro y simple. Era simplemente un reflejo de una asignación arbitraria de tratamiento bajo la hipótesis nula aguda de Fisher, y por casualidad, esta asignación generó un estadístico de prueba de .

En resumen: tenemos una asignación de tratamiento particular y su estadístico de prueba correspondiente. Si asumimos la hipótesis nula aguda de Fisher, ese estadístico de

prueba es simplemente una extracción de un proceso aleatorio. Y si esto es cierto, podemos barajar la asignación de tratamiento, calcular un nuevo estadístico de prueba y, finalmente, comparar este estadístico de prueba "falso" con el real.

La idea clave de la inferencia de aleatorización es que, bajo la hipótesis nula aguda, la asignación de tratamiento en última instancia no importa. Se asume explícitamente que, al pasar de una asignación a otra, los contrafácticos no cambian; siempre son iguales a los resultados observados. Por lo tanto, la distribución de aleatorización es simplemente un conjunto de todos los estadísticos de prueba posibles para cada vector de asignación de tratamiento posible. Los pasos tercero y cuarto amplían esta idea, barajando literalmente la asignación de tratamiento y calculando el estadístico de prueba único para cada asignación. Al repetir esto (paso 5), en el límite se recorrerán todas las combinaciones posibles que producirán una distribución de estadísticos de prueba bajo la hipótesis nula aguda.

Una vez que se tenga la distribución completa de los estadísticos de prueba, se puede calcular el valor exacto. ¿Cómo? Sencillo: se clasifican estos estadísticos de prueba, se ajusta el efecto real a esa clasificación, se cuenta el número de estadísticos de prueba falsos que predominan sobre el real y se divide ese número entre todas las combinaciones posibles. Formalmente, sería así:

$$\Pr\left(t(D',Y) \geq t(D_{observed},Y) \mid \delta_i = 0, \forall i)\right) = \frac{\sum_{D' \in \Omega} \quad I\left(t(D',Y) \geq t(D_{observed},Y)\right)}{K}$$

Si observa detenidamente la Tabla 15, verá que, para cada unidad, solo observamos un resultado potencial. Sin embargo, bajo la hipótesis nula aguda, podemos inferir el otro contrafáctico faltante. Solo disponemos de información sobre los resultados observados con base en la ecuación de conmutación. Por lo tanto, si una unidad recibe tratamiento, conocemos su Y1, pero no su Y0.

El segundo paso es la construcción de lo que se denomina un "estadístico de prueba". ¿Qué es esto? Un estadístico de prueba t(D,Y) es simplemente una cantidad escalar conocida que se calcula a partir de las asignaciones de tratamiento y los resultados observados. A menudo, no es más que una medida de la relación entre los valores de Y por D. En el resto de esta sección, desarrollaremos diversas maneras de construir estadísticos de prueba, pero comenzaremos con una medida bastante sencilla: la diferencia en el resultado medio.

En última instancia, los estadísticos de prueba nos ayudan a distinguir entre la hipótesis nula aguda en sí y alguna otra hipótesis. Y si se desea un estadístico de prueba con alta potencia estadística, se necesita que este adopte valores "extremos" (es decir, grandes en valores absolutos) cuando la hipótesis nula es falsa, y es necesario que estos valores grandes sean improbables cuando la hipótesis nula es verdadera.24

Como dijimos, existen varias maneras de estimar un estadístico de prueba, y analizaremos varias de ellas, pero comencemos con la simple diferencia en los

resultados medios. Los valores promedio para el grupo de tratamiento son 34/4, los valores promedio para el grupo de control son 30/4, y la diferencia entre estos dos promedios es 1. Por lo tanto, dada esta asignación de tratamiento particular en nuestra muestra (la asignación verdadera, cabe mencionar), existe un estadístico de prueba correspondiente (la simple diferencia en los resultados medios) que es igual a 1.

Ahora bien, lo que implica la hipótesis nula aguda de Fisher es uno de los aspectos más interesantes de este método. Si bien históricamente desconocemos el contrafactual de cada unidad, bajo la hipótesis nula aguda sí conocemos el contrafactual de cada unidad. ¿Cómo es posible? Porque si ninguna de las unidades tiene efectos de tratamiento distintos de cero, entonces cada contrafáctico debe ser igual a su resultado observado. Esto significa que podemos completar los contrafácticos faltantes con los valores observados (Tabla 4.7).

Table 4.7: Example of made-up data for 8 people with filled-in counterfactuals according to Fisher's sharp null hypothesis.

Name	D	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Con estos contrafácticos faltantes reemplazados por el resultado observado correspondiente, no hay efecto del tratamiento a nivel de unidad y, por lo tanto, un ATE de cero. Entonces, ¿por qué encontramos antes una simple diferencia en los resultados medios de 1 si, de hecho, no hubo efecto promedio del tratamiento? Simple: era solo ruido, puro y simple. Era simplemente un reflejo de una asignación arbitraria de tratamiento bajo la hipótesis nula aguda de Fisher, y por casualidad, esta asignación generó un estadístico de prueba de .

En resumen: tenemos una asignación de tratamiento particular y su estadístico de prueba correspondiente. Si asumimos la hipótesis nula aguda de Fisher, ese estadístico de prueba es simplemente una extracción de un proceso aleatorio. Y si esto es cierto, podemos barajar la asignación de tratamiento, calcular un nuevo estadístico de prueba y, finalmente, comparar este estadístico de prueba "falso" con el real.

La idea clave de la inferencia de aleatorización es que, bajo la hipótesis nula aguda, la asignación de tratamiento en última instancia no importa. Se asume explícitamente que, al pasar de una asignación a otra, los contrafácticos no cambian; siempre son iguales a los resultados observados. Por lo tanto, la distribución de aleatorización es simplemente un conjunto de todos los estadísticos de prueba posibles para cada vector de asignación de tratamiento posible. Los pasos tercero y cuarto amplían esta idea, barajando literalmente la asignación de tratamiento y calculando el estadístico de prueba único para cada asignación. Al repetir esto (paso 5), en el límite se recorrerán todas las combinaciones posibles que producirán una distribución de estadísticos de prueba bajo la hipótesis nula aguda.

Una vez que se tenga la distribución completa de los estadísticos de prueba, se puede calcular el valor exacto. ¿Cómo? Sencillo: se clasifican estos estadísticos de prueba, se ajusta el efecto real a esa clasificación, se cuenta el número de estadísticos de prueba falsos que predominan sobre el real y se divide ese número entre todas las combinaciones posibles. Formalmente, sería así:

$$\Pr\left(t(D', Y) \geq t(D_{observed}, Y) \mid \delta_i = 0, \forall i)\right) = \frac{\sum_{D' \in \Omega} \quad I\left(t(D', Y) \geq t(D_{observed}, Y)\right)}{K}$$

De nuevo, vemos qué significa "exacto". Estos valores p son exactos, no aproximaciones. Y con un umbral de rechazo de α (por ejemplo, 0,05), una prueba de inferencia de aleatorización rechazará falsamente la hipótesis nula aguda menos del 100 × α % de las veces.

4.2.4 Ejemplo

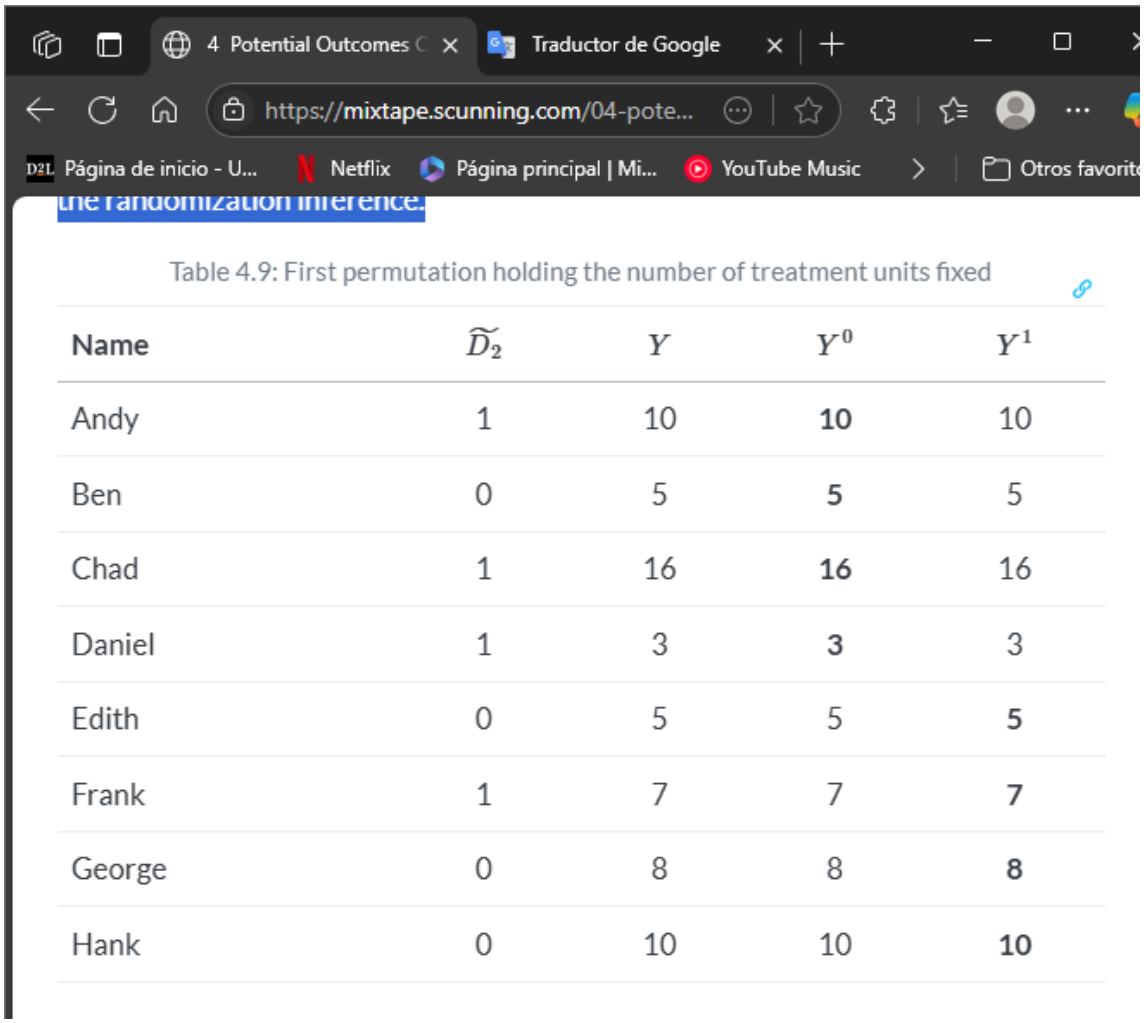
Creo que esto ha sido un poco abstracto, y cuando las cosas son abstractas, es fácil confundirse, así que analicemos un ejemplo con datos nuevos. Imaginemos que trabajamos en un albergue para personas sin hogar con un programa de terapia cognitivo-conductual (TCC) para el tratamiento de enfermedades mentales y abuso de sustancias. Disponemos de fondos suficientes para incluir a cuatro personas en el estudio, pero tenemos ocho residentes. Por lo tanto, hay cuatro en tratamiento y cuatro en control. Tras finalizar la TCC, se entrevista a los residentes para determinar la gravedad de sus síntomas de enfermedad mental. El terapeuta registra su salud mental en una escala del 0 al 20. Con la siguiente información, podemos completar los contrafácticos faltantes para satisfacer la hipótesis nula aguda de Fisher y calcular el estadístico de prueba correspondiente basado en esta asignación de tratamiento. Nuestro estadístico de prueba será el valor absoluto de la diferencia simple en los resultados medios para

simplificar. La estadística de prueba para esta asignación de tratamiento particular es simplemente |34/4−30/4|=8,5−7,5=1, utilizando los datos de la Tabla 4.8.

Table 4.8: Self-reported mental health for 8 residents in a homeless shelter (treatment and control).

Name	$D_1(\$15)$	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	1	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Ahora pasamos a la etapa de aleatorización. Barajemos la asignación del tratamiento y calculemos el nuevo estadístico de prueba para ese nuevo vector de tratamiento. La Tabla 4.9 muestra esta permutación. Pero primero, una cosa. Vamos a mantener el número de unidades de tratamiento fijo en este ejemplo. Pero si la asignación del tratamiento hubiera seguido algún proceso aleatorio, como el de Bernoulli, entonces el número de unidades de tratamiento sería aleatorio y la asignación aleatoria del tratamiento sería mayor que la que estamos haciendo aquí. ¿Cuál es correcto? Ninguna es correcta en sí misma. Mantener las unidades de tratamiento fijas es, en última instancia, un reflejo de si se habían fijado en la asignación del tratamiento original. Esto significa que necesita conocer sus datos y el proceso mediante el cual se asignaron las unidades al tratamiento para saber cómo realizar la inferencia de aleatorización.



the randomization inference.

Table 4.9: First permutation holding the number of treatment units fixed

Name	$\widetilde{D_2}$	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	0	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	1	7	7	**7**
George	0	8	8	**8**
Hank	0	10	10	**10**

Con esta reorganización de la asignación de tratamiento, podemos calcular un nuevo estadístico de prueba, que es |36/4−28/4|=9−7=2. Antes de continuar, observemos este estadístico de prueba: el estadístico de prueba de 2 es falso porque no corresponde a la verdadera asignación de tratamiento. Sin embargo, bajo la hipótesis nula, la asignación de tratamiento ya carecía de sentido, ya que no existían efectos de tratamiento distintos de cero. La cuestión es que, incluso cuando se cumple la hipótesis nula de ausencia de efecto, puede, y normalmente lo hará, producir un efecto distinto de cero simplemente por las propiedades de la muestra finita.

Anotemos ese número 2 y realicemos otra permutación; es decir, reorganizamos de nuevo la asignación de tratamiento. La Tabla 4.10 muestra esta segunda permutación, manteniendo de nuevo el número de unidades de tratamiento fijo en cuatro en el grupo de tratamiento y cuatro en el grupo de control.

Name	$\widetilde{D_2}$	Y	Y^0	Y^1
Andy	1	10	**10**	10
Ben	0	5	**5**	5
Chad	1	16	**16**	16
Daniel	1	3	**3**	3
Edith	0	5	5	**5**
Frank	0	7	7	**7**
George	1	8	8	**8**
Hank	0	10	10	**10**

El estadístico de prueba asociado con esta asignación de tratamiento es |36/4−27/4|=9−6,75=2,25. Nuevamente, 2,25 es un valor extraído de una asignación de tratamiento aleatoria donde cada unidad no tiene efecto.

Cada vez que se aleatoriza la asignación de tratamiento, se calcula un estadístico de prueba, se almacena en algún lugar y luego se pasa a la siguiente combinación. Se repite esto una y otra vez hasta agotar todas las posibles asignaciones de tratamiento. Veamos las primeras iteraciones en la Tabla 4.11.

Assignment	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	$\lvert T_i \rvert$
True D	1	1	1	1	0	0	0	0	1
$\widetilde{D_2}$	1	0	1	1	0	1	0	0	2
$\widetilde{D_3}$	1	0	1	1	0	0	1	0	2.25

El paso final es calcular el valor exacto. Para ello, tenemos un par de opciones. Podemos usar software, que es una buena opción, o hacerlo manualmente. Por razones pedagógicas, prefiero hacerlo manualmente. Así que, empecemos.

Codigo stata

use https://github.com/scunning1975/mixtape/raw/master/ri.dta, clear

```
tempfile ri
gen id = _n
save "`ri'", replace

* Create combinations
* ssc install percom
combin id, k(4)
gen permutation = _n
tempfile combo
save "`combo'", replace

forvalue i =1/4 {
   ren id_`i' treated`i'
}



destring treated*, replace
cross using `ri'
sort permutation name
replace d = 1 if id == treated1 | id == treated2 | id == treated3 | id == treated4
replace d = 0 if ~(id == treated1 | id == treated2 | id == treated3 | id == treated4)

* Calculate true effect using absolute value of SDO
egen    te1 = mean(y) if d==1, by(permutation)
egen    te0 = mean(y) if d==0, by(permutation)

collapse (mean) te1 te0, by(permutation)
gen     ate = te1 - te0
keep    ate permutation

gsort -ate
```

```
gen rank = _n

su rank if permutation==1

gen pvalue = (` r(mean)'/70)

list pvalue if permutation==1

* pvalue equals 0.6
```

Este programa fue bastante sencillo debido a la pequeña cantidad de combinaciones posibles. De ocho observaciones, cuatro eligen ocho, lo que equivale a 70. Solo tuvimos que manipular los datos para llegar a ese punto, pero una vez que lo hicimos, el cálculo real fue sencillo. Por lo tanto, podemos ver que el ATE estimado no puede rechazar la hipótesis nula en la distribución placebo.

Sin embargo, a menudo los conjuntos de datos con los que trabajamos son mucho mayores que ocho observaciones. En esas situaciones, no podemos usar este método, ya que el volumen de combinaciones crece rápidamente a medida que n aumenta. Esperaremos por ahora la revisión de este método de inferencia cuando n sea demasiado grande hasta que hayamos tenido la oportunidad de profundizar en el tema.

4.2.5 Otros estadísticos de prueba

Recuerde que el segundo paso de esta metodología fue la selección del estadístico de prueba.25 Se seleccionó la diferencia simple en los resultados medios (o su valor absoluto), lo cual es adecuado cuando los efectos son aditivos y hay pocos valores atípicos en los datos. Sin embargo, los valores atípicos crean problemas para dicho estadístico de prueba debido a la variación que se introduce en la distribución de aleatorización. Por lo tanto, otros estadísticos de prueba alternativos resultan más atractivos.

Una transformación que maneja los valores atípicos y la asimetría de forma más general es la transformación logarítmica. Imbens y Rubin (2015) la definen como la diferencia promedio en una escala logarítmica por estado de tratamiento, o

$$T_{\log} = |\frac{1}{N_T}\sum_{i=1}^{N} \quad D_i \ln(Y_i) - \frac{1}{N_C}\sum_{i=1}^{N} \quad (1-D_i)\ln(Y_i)|$$

Esto tiene sentido cuando los datos brutos están sesgados, lo que ocurre con valores positivos como los ingresos y en casos en que los efectos del tratamiento son multiplicativos en lugar de aditivos.

Otro estadístico de prueba observado es el valor absoluto en la diferencia de cuantiles. Esto también protege contra valores atípicos y se representa como

$$T_{\text{median}} = |\text{median}(Y_T) - \text{median}(Y_C)|$$

Podríamos observar la mediana, el cuartil 25, el cuartil 75 o cualquier valor a lo largo del intervalo unitario.

El problema de los valores atípicos también nos lleva a considerar un estadístico de prueba que utiliza rangos en lugar de diferencias. Esto, de nuevo, resulta útil cuando hay un gran número de valores atípicos, cuando los resultados son continuos o los conjuntos de datos son pequeños. Los estadísticos de rango transforman los resultados en rangos y luego realizan un análisis sobre los propios rangos. La idea básica es clasificar los resultados y luego comparar el rango promedio de los grupos de tratamiento y control. Ilustremos esto primero con un ejemplo (Tabla 4.12).

Name	D	Y	Y^0	Y^1	Rank	R_i
Andy	1	10	**10**	10	6.5	2
Ben	1	5	**5**	5	2.5	-2
Chad	1	16	**16**	16	8	3.5
Daniel	1	3	**3**	3	1	-3.5
Edith	0	5	5	**5**	2.5	-1
Frank	0	7	7	**7**	4	-0.5
George	0	8	8	**8**	5	0.5
Hank	0	10	10	**10**	6.5	2

Como antes, solo observamos la mitad de los resultados potenciales dada la ecuación de conmutación que asigna resultados potenciales a resultados reales. Sin embargo, bajo la hipótesis nula aguda de Fisher, podemos imputar el contrafáctico faltante para garantizar que no haya efecto del tratamiento. Para calcular los rangos, simplemente contamos el número de unidades con valores mayores de Y, incluida la unidad en cuestión. Y en caso de empate, simplemente tomamos el promedio de todas las unidades empatadas.

Por ejemplo, considere a Andy. Andy tiene un valor de 10. Andy es tan grande como él mismo (1); más grande que Ben (2), Daniel (3), Edith (4), Frank (5) y George (6); y empatado con Hank (7). Como está empatado con Hank, promediamos los dos, lo que eleva su rango a 6,5. Ahora considere a Ben. Ben tiene un valor de 5. Es tan grande como él mismo (1), más grande que Daniel (2) y empatado con Edith (3). Por lo tanto, promediamos a Edith y a él mismo para obtener 0,5, lo que nos lleva a un rango de 2.

Sin embargo, es común normalizar los rangos para obtener una media de 0, lo cual se realiza según la siguiente fórmula:

$$\tilde{R}_i = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} I\left(Y_j \leq Y_i\right) - \frac{N+1}{2}$$

Esto nos da la columna final, que ahora usaremos para calcular el estadístico de prueba. Usemos el valor absoluto de la diferencia simple en los resultados promedio en el rango normalizado, que aquí es

$$T_{\text{rank}} = \left|0 - \frac{1}{4}\right| = \frac{1}{4}$$

Para calcular el valor p exacto, simplemente realizaríamos el mismo proceso de aleatorización que antes, solo que en lugar de calcular la diferencia simple en los resultados medios, calcularíamos el valor absoluto de la diferencia más simple en el rango medio.

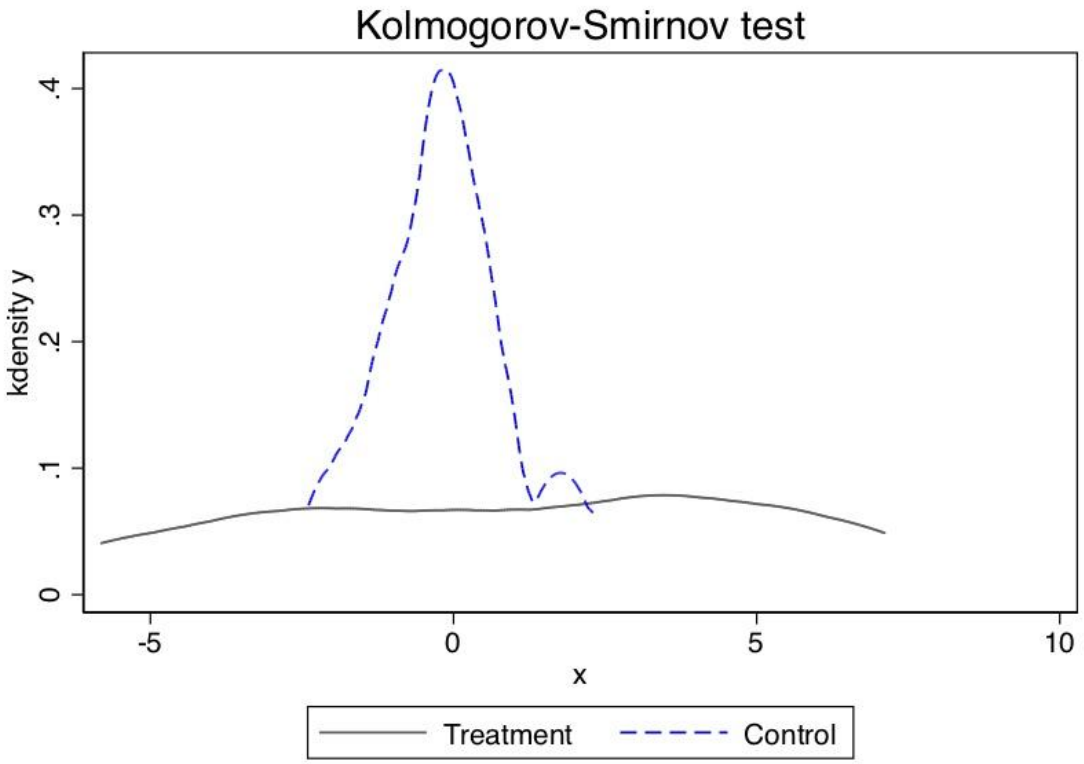
Sin embargo, todos estos estadísticos de prueba que hemos estado analizando han sido diferencias en los resultados según el estado de tratamiento. Consideramos diferencias simples en promedios, diferencias simples en promedios logarítmicos, diferencias en cuantiles y diferencias en rangos. Imbens y Rubin (2015) señalan que existen deficiencias derivadas de centrarse únicamente en algunas características de los datos (p. ej., la asimetría), ya que puede hacer que pasemos por alto diferencias en otros aspectos. Esto, en particular, puede ser problemático si la varianza en los resultados potenciales del grupo de tratamiento difiere de la del grupo de control. Centrarse únicamente en las diferencias de promedios simples que analizamos podría no generar valores p lo suficientemente extremos como para rechazar la hipótesis nula, incluso cuando esta no se cumple. Por lo tanto, podría interesarnos un estadístico de prueba que pueda detectar diferencias en las distribuciones entre las unidades de tratamiento y control. Una de estas pruebas estadísticas es la prueba de Kolmogorov-Smirnov.

Definamos primero la función de distribución acumulativa (FDA) empírica como:

$$\begin{aligned} \hat{F}_C(Y) &= \frac{1}{N_C} \sum_{i:D_i=0} 1(Y_i \leq Y) \\ \hat{F}_T(Y) &= \frac{1}{N_T} \sum_{i:D_i=1} 1(Y_i \leq Y) \end{aligned}$$

Si dos distribuciones son iguales, su CDF empírica también lo es. Sin embargo, cabe destacar que las CDF empíricas son funciones y los estadísticos de prueba son escalares. Entonces, ¿cómo convertiremos las diferencias entre dos funciones en una única cantidad escalar? Fácil: usaremos la diferencia máxima entre las dos CDF empíricas. Visualmente, será literalmente la mayor distancia vertical entre las dos CDF empíricas. Esa distancia vertical será nuestro estadístico de prueba. Formalmente, es:

$$T_{KS} = max|\hat{F}_T(Y_i) - \hat{F}_C(Y_i)|$$




```
clear
input  d y
0   0.22
0  -0.87
0  -2.39
0  -1.79
0   0.37
0  -1.54
0   1.28
0  -0.31
0  -0.74
0   1.72
0   0.38
0  -0.17
0  -0.62
0  -1.10
```

0 0.30

0 0.15

0 2.30

0 0.19

0 -0.50

0 -0.09

1 -5.13

1 -2.19

1 -2.43

1 -3.83

1 0.50

1 -3.25

1 4.32

1 1.63

1 5.18

1 -0.43

1 7.11

1 4.87

1 -3.10

1 -5.81

1 3.76

1 6.31

1 2.58

1 0.07

1 5.76

1 3.50

end

twoway (kdensity y if d==1) (kdensity y if d==0, lcolor(blue) lwidth(medium) lpattern(dash)), \\\

title(Kolmogorov-Smirnov test) legend(order(1 ``Treatment'' 2 ``Control''))

Para calcular el valor p, se repite lo que hicimos en los ejemplos anteriores. En concreto, se elimina la variable de tratamiento, se reordenan los datos, se reasignan nuevos valores de tratamiento (fijos), se calcula el TKS, se guarda el coeficiente y se repite el proceso mil veces o más hasta obtener una distribución que permita calcular un valor p empírico.

4.2.6 Inferencia de aleatorización con n grande

¿Qué hicimos cuando el número de observaciones es muy grande? Por ejemplo, la muestra total de Thornton fue de 2901 participantes. De ellos, 2222 recibieron algún incentivo. Wolfram Alpha es una calculadora en línea fácil de usar para cálculos más complejos y una interfaz intuitiva. Si ingresa al sitio web y escribe "2901 choose 2222", obtendrá el siguiente número truncado de combinaciones:

615056610949825151369928033330771847162379504341926926182640318266385758921095807995693142554352679783785174154933743845244511660523651518050517786402824289794087767092848717201188223218885942515735991356144283120935017438277464692155849858790123688111563011540267646207996405072248645607065160780040934113065544540016312151177000750339179099962167196885539725968603122868768036473093648093307466530 7...

Buena suerte calculando esas combinaciones. Así que, claramente, los valores p exactos utilizando todas las combinaciones no funcionarán. Por lo tanto, estimaremos valores p aproximados. Para ello, necesitaremos asignar aleatoriamente el tratamiento, estimar un estadístico de prueba que satisfaga la hipótesis nula aguda para esa muestra, repetirlo miles de veces, y luego calcular el valor p asociado con la asignación de este tratamiento según su posición en la distribución.

Codigo STATA

```
use https://github.com/scunning1975/mixtape/raw/master/thornton_hiv.dta, clear

tempfile hiv
save "`hiv'", replace

* Calculate true effect using absolute value of SDO
egen    te1 = mean(got) if any==1
egen    te0 = mean(got) if any==0

collapse (mean) te1 te0
gen     ate = te1 - te0
```

```
keep    ate
gen iteration = 1

tempfile permute1
save "`permute1'", replace

* Create a hundred datasets

forvalues i = 2/1000 {

use "`hiv'", replace

drop any
set seed `i'
gen random_`i' = runiform()
sort random_`i'
gen one=_n
drop random*
sort one

gen     any = 0
replace any = 1 in 1/2222

* Calculate test statistic using absolute value of SDO
egen    te1 = mean(got) if any==1
egen    te0 = mean(got) if any==0

collapse (mean) te1 te0
gen     ate = te1 - te0
keep    ate
```

```
gen    iteration = `i'
tempfile permute`i'
save "`permute`i'", replace

}

use "`permute1'", replace
forvalues i = 2/1000 {
   append using "`permute`i'"
}

tempfile final
save "`final'", replace

* Calculate exact p-value
gsort -ate
gen rank = _n
su rank if iteration==1
gen pvalue = rank/1000
list if iteration==1
```

Table 4.13: Estimated p-value using different number of trials.

ATE	Iteration	Rank	p	No. trials
0.45	1	1	0.01	100
0.45	1	1	0.002	500
0.45	1	1	0.001	1000

Impresionante. La Tabla 4.13 muestra que el experimento de Thornton, bajo la hipótesis nula aguda de Fisher, con entre 100 y 1000 extracciones repetidas, arroja valores p altamente significativos. De hecho, siempre es el ATE de mayor rango en una prueba unilateral.

Así pues, lo que he hecho aquí es obtener una aproximación del valor p asociado con nuestro estadístico de prueba y la hipótesis nula aguda. En la práctica, si el número de extracciones es elevado, el valor p basado en esta muestra aleatoria será bastante preciso (Imbens y Rubin, 2015). Quería ilustrar este método de aleatorización porque, en realidad, es exactamente lo que se hará la mayor parte del tiempo, ya que el número de combinaciones con cualquier conjunto de datos de tamaño razonable será computacionalmente prohibitivo.

Ahora bien, en cierto modo, este ejercicio de aleatorización no reveló mucho, y probablemente se deba a que los hallazgos originales de Thornton eran demasiado precisos desde el principio (0,4 con un error estándar de 0,02). Podríamos lanzar bombas atómicas contra este resultado y no funcionaría. Pero el propósito principal es demostrar su robustez bajo diferentes maneras de generar esos valiosos valores p, así como proporcionar un mapa para programarlo usted mismo y para tener una forma intuitiva, posiblemente independiente, de pensar en la significancia misma.

4.2.7 Apalancamiento

Antes de concluir, me gustaría retomar algo que mencioné anteriormente sobre el apalancamiento. Un reciente y estimulante estudio de Young (2019) nos ha alertado sobre los desafíos que podemos enfrentar al usar la inferencia tradicional para estimar la incertidumbre de alguna estimación puntual, como los errores estándar robustos. Young encuentra problemas prácticos con nuestras formas tradicionales de inferencia, que, si bien se conocían previamente, no se habían destacado tanto como en su estudio. El problema que destaca es el del apalancamiento concentrado. El apalancamiento es una medida del grado en que una sola observación en la variable del lado derecho adquiere valores extremos e influye en la estimación de la pendiente de la línea de regresión. Una concentración de influencia incluso en unas pocas observaciones puede hacer que los coeficientes y los errores estándar sean extremadamente volátiles e incluso sesguen errores estándar robustos hacia cero, lo que conlleva tasas de rechazo más altas.

Para ilustrar este problema, Young (2019) realizó un ejercicio sencillo. Recopiló más de cincuenta artículos experimentales (de laboratorio y de campo) de las revistas insignia de la Asociación Económica Americana: American Economic Review, American Economic Journal: Applied y American Economic Journal: Economic Policy. Posteriormente, volvió a analizar estos artículos, utilizando los modelos de los autores, eliminando una observación o conglomerado y reestimando el modelo completo repetidamente. Su hallazgo fue impactante:

Con la eliminación de una sola observación, el 35 % de los resultados reportados con una significación de 0,01 en el artículo promedio pueden volverse insignificantes a ese nivel. Por el contrario, el 16 % de los resultados reportados con una significación de 0,01 pueden considerarse significativos a ese nivel. (567)

Que la evidencia dependa tanto de unas pocas observaciones genera dudas sobre la claridad de nuestro trabajo, así que ¿cuáles son nuestras alternativas? El método de inferencia aleatoria basado en la hipótesis nula aguda de Fisher, que se analizará en esta

sección, puede mejorar estos problemas de apalancamiento, además de las razones mencionadas para considerarlo. En un artículo típico, la inferencia aleatoria encontró efectos de tratamiento individuales que representaron entre un 13 % y un 22 % menos de resultados significativos que los descubiertos por el propio análisis de los autores. Al parecer, la inferencia aleatoria es algo más robusta ante la presencia de apalancamiento en algunas observaciones.

4.3 Conclusión

En conclusión, en este capítulo hemos realizado algunas cosas. Hemos introducido la notación de resultados potenciales y la hemos utilizado para definir varios tipos de efectos causales. Demostramos que la diferencia simple en los resultados medios era igual a la suma del efecto medio del tratamiento, o sesgo de selección, y el sesgo ponderado del efecto del tratamiento heterogéneo. Por lo tanto, el estimador simple de diferencia en los resultados medios está sesgado a menos que los términos segundo y tercero se anulen. Una situación en la que se anulan es bajo independencia del tratamiento, es decir, cuando el tratamiento se ha asignado independientemente de los resultados potenciales. ¿Cuándo se produce la independencia? La situación más frecuente se da durante la aleatorización física del tratamiento a las unidades. Dado que la aleatorización física asigna el tratamiento por razones independientes de los resultados potenciales, el sesgo de selección se anula, al igual que el sesgo del efecto heterogéneo del tratamiento. A continuación, analizamos una segunda situación en la que ambos términos se anulan: la independencia condicional.

CONCLUSIONES

1. Esta breve historia se centrará en el desarrollo del modelo de resultados potenciales. Véase Morgan (1991) para una historia más completa de las ideas econométricas.↩

2. Alrededor de los 20 años, finalmente terminé Tomb Raider 2 en la Sony PlayStation. Así que sí, me identifico totalmente con los logros de Gauss a tan temprana edad.↩

3. Para más información sobre la transición de Splawa-Neyman (1923) a Roland A. Fisher (1925), véase D. B. Rubin (2005).↩

4. En el grupo placebo, se inyectó una solución salina a los niños.↩

5. Puede encontrar más información sobre este fascinante experimento en Newhouse (1993).↩

6. Si fuera un apostador —y lo soy—, apostaría a que se entregará al menos un premio experimental más. El candidato más probable es John List, por su trabajo con experimentos de campo.↩

7. Curiosamente, la filosofía, como campo, emprende una cuidadosa consideración de los contrafácticos al mismo tiempo que los primeros trabajos de Rubin con el gran filósofo metafísico David Lewis (Lewis, 1973). Aparentemente, esta información estaba en el aire, lo que dificulta rastrear el efecto causal de las ideas científicas.↩

8. El razonamiento contrafáctico puede ser útil, pero también perjudicial, sobre todo cuando genera arrepentimiento. Probablemente exista una versión contrafáctica de la falacia del coste hundido: dado que no podemos saber con certeza qué habría sucedido si hubiéramos tomado una decisión diferente, debemos aceptar cierta incertidumbre

básica solo para superarla. En última instancia, nadie puede afirmar que una decisión alternativa habría tenido un mejor resultado. No se puede saber, y eso a veces puede ser difícil. Lo ha sido y seguirá siendo, al menos para mí.↩

9. Hasta donde sé, el filósofo que mencioné antes, David Lewis, creía que los resultados potenciales eran en realidad mundos separados, tan reales como nuestro mundo. Eso significa que, según Lewis, existe un mundo muy real, aunque inaccesible, en el que Kanye liberó a Yandhi en lugar de Jesús es Rey, lo cual me resulta extremadamente frustrante.↩

10. Un par de cosas. En primer lugar, este análisis puede extenderse a más de dos resultados potenciales, pero como gran parte de este libro se centra en la evaluación de programas, me centraré solo en dos. En segundo lugar, el tratamiento aquí es cualquier intervención particular que pueda manipularse, como tomar aspirina o no. En la tradición de los resultados potenciales, la manipulación es fundamental para el concepto de causalidad.↩

11. Esto puede ocurrir debido a las preferencias, pero también debido a las limitaciones. La maximización de la utilidad, recuerde, es un proceso de optimización restringido y, por lo tanto, tanto el valor como los obstáculos influyen en la clasificación.↩

12. Piense en el "médico perfecto" como un Sombrero Seleccionador al estilo de Harry Potter. Aprendí por primera vez esta ilustración del "médico perfecto" del propio Rubin.↩

13. La razón por la que la UTA es negativa se debe a que el tratamiento en este caso es la cirugía, que no tuvo el mismo éxito que las unidades no tratadas con quimioterapia. Pero esto podría interpretarse fácilmente como 3,2 años adicionales de vida si hubieran recibido quimioterapia en lugar de cirugía.↩

14. $ATE = p \times ATT + (1-p) \times ATU = 0{,}5 \times 4{,}4 + 0{,}5 \times -3{,}2 = 0{,}6$.↩

15. Nótese que Angrist y Pischke (2009) tienen una descomposición ligeramente diferente donde el SDO=ATT+sesgo de selección, pero eso se debe a que su parámetro de interés es el ATT y, por lo tanto, el tercer término no aparece.↩

16. ¿Por qué digo "ganancias"? Porque la ganancia de la cirugía es $Yi1 - Yi0$. Por lo tanto, si decimos que es independiente de las ganancias, estamos diciendo que es independiente de Y1 e Y0.↩

17. Aquí es donde la economía es realmente útil en mi opinión. La economía enfatiza que los valores observados son equilibrios basados en agentes que participan en una optimización restringida y que prácticamente garantiza que la independencia se viole en los datos observacionales. Rara vez los seres humanos toman decisiones importantes en la vida lanzando monedas.↩

18. En Craig (2006), un banquero que jugaba al póquer usó el segundero de su reloj como generador de números aleatorios para farolear aleatoriamente cuando tenía una mano débil.↩

19. Debido a que no está preclasificado, al ejecutarlo, su respuesta será cercana pero ligeramente diferente debido a la aleatoriedad de la muestra extraída.↩

20. He aquí una forma sencilla de recordar qué igualdad obtenemos con la independencia. El término antes de la barra vertical es el mismo, pero el término después de la barra vertical es diferente. Por lo tanto, la independencia garantiza que, en la población, Y1 sea el mismo en promedio para cada grupo.↩

21. También optó por agrupar esos errores estándar por aldea para 119 aldeas. Al hacerlo, aborda el problema del rechazo excesivo que vimos anteriormente al hablar de la agrupación en el capítulo de probabilidad y regresión.↩

22. Aparentemente, Bristol acertó las cuatro tazas de té.↩

23. Simplemente quiero decir que la inferencia no depende de asintóticas ni de ningún tipo de distribución en el proceso de generación de datos.↩

24. Resulta interesante comprender el mecanismo preciso de este método: en realidad, no está diseñado para detectar pequeños efectos del tratamiento, ya que a menudo estos pequeños valores se ven superados por el proceso de aleatorización. Sin embargo, no hay ninguna razón filosófica para creer que los efectos promedio del tratamiento deban ser relativamente grandes. Simplemente, la inferencia de aleatorización sí lo requiere para distinguir el efecto real del de la hipótesis nula aguda.↩

25. Para un análisis más profundo de los siguientes temas, recomiendo encarecidamente el excelente Imbens y Rubin (2015), en particular el capítulo 5.↩